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

A Feature Level Fusion Approach for Object Classification

Stefan Wender, Klaus C. J. Dietmayer

***Abstract*— A new feature level fusion approach for object classification is introduced. The system is implemented to fuse sensor data of a laser scanner and a video sensor. A new method of video feature extraction incorporates features, which are obtained from the laser scanner, to handle the problem of multiple views of cars. The laser scanner's estimates of contour information can identify the discrete sides of rectangular objects. These object sides are transformed to the video image. A perspective reconstruction compensates deformations as well as size differences in the video image. Afterwards, an object detector is applied. A new method performs a feature extraction from this detector. The classification algorithms fuse these new features with additional features, which are obtained from the laser scanner and the tracking algorithms. The complete system is applicable in real time. An evaluation with labeled real world test data is given.**

I. Introduction

ADVANCED Driver Assistant Systems (ADAS) are in the focus of today's research and development. Although recent ADAS usually rely on an exclusively used sensor, future systems can benefit from a shared sensor platform. For this purpose, a common data processing unit is needed, which can fuse different sensors and provide the preprocessed information in terms of a vehicle environment model to several applications. This environment model should contain a list of observed objects, coupled with information like dynamic state and class.

A laser scanner based approach for object detection, tracking, and classification was proposed in former works [1]. However, the appropriate usage of additional sensor information usually can significantly improve the performance of such a system. The successful application of a feature based fusion of laser scanner and video data for the purpose of tracking was already demonstrated [2].

A new feature level fusion approach for the purpose of object classification extends these tracking algorithms. This approach is implemented and analyzed for a laser scanner and a video camera. Five classes are distinguished by the classification: "Pedestrian", "Bike", "Car", "Truck", and the remaining objects of the class "Unknown".

The primary objective of this fusion approach is the improvement of the laser scanner's classification performance for non-moving objects. While the laser scanner based framework already showed good results in the classification performance, several objects are still misclassified. One reason of misclassifications consists of the relatively high influence of the estimated object velocity on the classification result. Objects are often misclassified, if they are not moving or due to missing velocity information during the first time steps of the Kalman Filter based tracking. The aim is to improve the classification of these objects, if they are in the field of view of the video camera.

Manuscript received January 15, 2007.

S. Wender is with the Institute of Measurement, Control and Microtechnology at University of Ulm, Albert-Einstein-Allee 41, 89081 Ulm, Germany (e-mail: stefan.wender@uni-ulm.de).

K. C. J. Dietmayer is Professor at the Institute of Measurement, Control and Microtechnology at University of Ulm

II. System Overview

A general layout is introduced, which can fuse multiple sensors at the feature level (Figure 1). Based on the available sensors, several features for the purpose of tracking and classification are extracted.

A tracking is performed by Kalman Filter estimation. A box object model is applied, which enables the estimation of position, size, and velocity [3]. Since improvements of the tracking by feature level fusion are discussed in [2], the following sections will concentrate on a feature level fusion approach to improve the object classification.

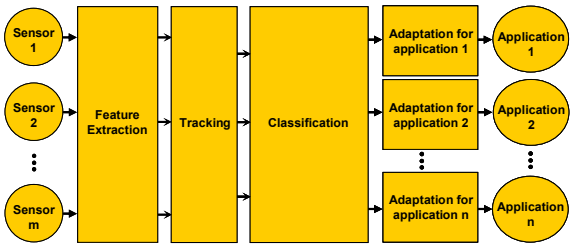

Fig. 1. System layout: Features are extracted from several sensors as well as from the tracking algorithms. The classification algorithms fuse all features and estimate membership values for each of the distinguished classes. The classification output and the selected object class can be optimized to different requirements of different applications simultaneously.

Based on the tracking, additional features are extracted to support the classification. The classification combines statistical and rule based approaches and calculates membership values for each of the distinguished classes. The output of the framework can be adapted to different requirements of different applications simultaneously.

III. Sensor Setup

The introduced fusion framework is implemented for evaluation purposes. This work uses an IBEO ALASCA XT laser scanner and a PCO Pixelfly video camera. In addition, it is possible to integrate DGPS and precise digital maps [4].

While the laser scanner's field of view (FOV) is very large, the field of view of the video camera is rather small (Figure 2). A wide angle lens is used for applications in urban scenarios, while a telephoto lens is recommended for highway scenarios. The use of the video camera is nevertheless promising, because the camera's FOV covers directly the area that the test vehicle will pass. It is expected, that the video camera can improve especially the classification performance for objects with low or zero velocity (non-moving cars, pedestrians, bikes, or trucks).

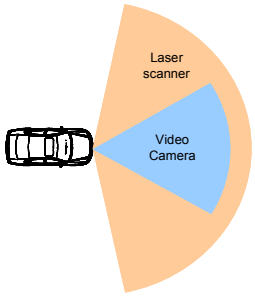

Fig. 2. Field of views (FOVs) of the used sensors: The FOV of the laser scanner is very large. The FOV of the video camera is rather small. Nevertheless, the camera's FOV covers the area in front of the vehicle, which is the most important area for ADAS. The detailed video measurements are expected to improve the classification in this area.

IV. Feature Extraction

Several types of features are calculated for the purpose of object classification. Some features like detected reflectors or contour and size information are directly calculated based on the laser scanner measurements. Other features like velocity are obtained from the tracking algorithms. Details about these two groups of features can be found in [5].

The last group of features is calculated by the application of pattern recognition algorithms on the video image. These features are described in the following section.

V. Video Features

A. Object Detection with a Boosted Cascade

The task of feature extraction from video images is performed by a complete object detection system. The cascaded classifier, which was developed by Viola et al. [6], is known from literature. The system is based on Haar like features. Appropriate feature combinations are selected by AdaBoost. Several combinations are used in a classifier cascade. Each stage of the cascade discards some of the object hypotheses. The remaining objects, which pass the last stage, are usually used as detected objects (Figure 3).

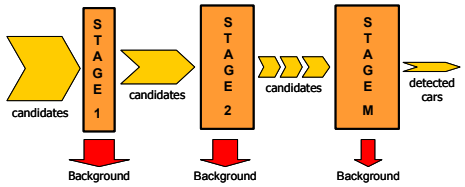

Fig. 3. Classifier cascade: Each cascade stage rejects some of the object candidates. The hypotheses, which pass the last stage, are usually used as detected objects.

This classifier was initially developed for face detection, but due to its performance it was already successfully applied to a wide range of objects (i.e. pedestrians [7], [8], heads [9], rear ends of cars [7], [10], [11]).

The detection of cars without restrictions to the viewing angle seems to be challenging, because a car's view significantly changes with its orientation to the video sensor. In addition, cars are the most frequent object class besides the class unknown, which represents the background. Therefore, the feature level fusion approach will be demonstrated for this class.

Features to detect pedestrians can be obtained from an additional cascade. First experiments already showed promising results, but the complete evaluation for pedestrian was not finished yet. Bike features may be calculated in a similar manner to car features. A feature extraction for trucks in the proposed manner seems only to be useful, if the trucks are completely in the field of view of the camera.

B. Challenges

There are several challenges corresponding to the car detection task. If only a single video sensor is used for object detection, there will be a lot of object hypotheses, which must be evaluated. Objects must be expected at all possible positions and in all possible sizes in the video image, if no additional knowledge is available. This is usually performed with image pyramids. The complete image scanning procedure then needs a lot of processing time. There are approaches, which reduce the available object hypotheses by a flat world assumption [11]. This

approach benefits from the correlation of object position and size. Unfortunately the flat world assumption does not hold for traffic scenarios on bumpy or hilly roads.

A more promising approach is the sensor fusion of distance sensors and video cameras. The distance sensors measure obstacles. The corresponding positions in the video image can be calculated, if the sensors are synchronous and calibrated. This approach enormously reduces the possible object size and position. Present works only concentrated on rear views of vehicles [2], [10].

Usually, the orientation of observed vehicles is only restricted on highways. At intersections and in urban areas, vehicles can occur in all possible orientations to the video sensor. The vehicle's appearance changes with its orientation to the video sensor (Figure 4).

Usually, pattern recognition concentrates on contour and texture information. For this reason the detection task becomes more complex.

Fig. 4. The appearance of cars significantly changes with the orientation to the video sensor. For this reason, the car detection task becomes more complex, because video object detection is usually based on texture and contour information.

This fact is considered by Schneiderman and Kanade [12]. They trained a detector with samples of different viewpoints. Thus, a complex detector, which can represent all views, was necessary. Another possibility is to train several detectors to cover different viewpoints with different detectors. Consequently many detectors have to be applied to the same image, since each detector can only cover small changes of the orientation.

Both solutions lead to quite high computational costs. Therefore, this work extends the idea of hypotheses selection by additionally obtaining the object's orientation from a laser scanner. This orientation allows for a reconstruction of a perspectively warped side of a vehicle (e.g. rear end, left side, right side, or front). Thus, only four different patterns have to be detected. For this purpose, the earlier mentioned pattern recognition system based on contour information and texture can be used. Experiments have shown that it is possible to train one detector for combined recognition of left and right side and one for front and rear ends, respectively.

C. *View Decomposition*

The laser scanner and the video camera work synchronously and are calibrated. Therefore, the positions and the orientations of the sensors relative to the vehicle are known. Usually, a sensor provides its measurements in its own coordinate system.

The calibration enables a transformation of a point from the laser scanner coordinate system to the image plane by applying rotational and translational matrices and the pinhole camera model [2]. An example of transformed laser scanner measurements is shown in Figure 5.

Fig. 5. Coordinate transformation: The measurements of the laser scanner as well as corresponding object features can be transformed to the video coordinate system due to calibrated and synchronous sensors. The example shows the transformed laser scanner measurements in the image.

The three dimensional shape of cars is approximated by a cuboid (Figure 6, left). This three dimensional object box is localized for all objects by the sensor information of the laser scanner.

The laser scanner estimates contour information. The calculated features provide information about the object distance, orientation and a visible corner (Figure 5, middle).

The video sensor provides a mixture of the visible object sides (Figure 6, right). The influence of a car roof on this mixture will be quite small due to the camera position. For this reason, the upper side will be ignored by further algorithms.

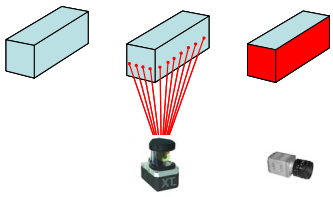

Fig. 6. The box as three dimensional object model (left), the contour information, measured by the laser scanner (middle) and the two object sides, which mainly influence the appearance of the object in the video data (right).

The laser scanner measurements of an object provide information about the vehicle's horizontal position and orientation. The laser scanner usually measures one or two sides of a cuboid. Consequently, a 3D object box can be fitted into the measurements. The fixed value of the cuboid's width and length is defined by the maximum length of a car. The vertical dimension of the box is quite uncertain, but it is possible to calculate an upper and a lower bound. The upper bound is given by the sum of the

laser scanner's lowest measurement and the maximum height of a car. The lower bound is given by the highest laser scanner measurement and the maximum car height.

The cuboid is transformed to the video coordinate system. Several features, which describe the size and position of the cuboid in the video coordinate system, are handed over to the video feature extraction.

The box side with the best visible orientation to the camera is used as the region of interest (ROI) for the object detection. This ROI is increased by 15 percent to ensure that it also contains some background, which is necessary for the object detection algorithms. An example of such a ROI is given in Figure 7.

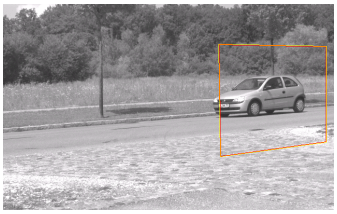

Fig. 7. Video ROI: An appropriate side of the object cuboid, which is estimated by the laser scanner, is transformed to the video image. The side is increased to contain some amount of background pixels around the object.

As the transformed box side describes the position and deformation of the original object side in the video image, a perspective warping can reconstruct the original view of the object side. The ROI has a fixed length in m in world coordinates. The warping creates a rectangular image of a fixed size in pixel. This process performs a compensation of distance based object size differences. For this reason, the later applied pattern recognition system can benefit from the a priori knowledge about the expected reconstructed object size. Only a small range of positions and sizes have to be evaluated by the detector. The reconstructed object side view is shown in Figure 8.

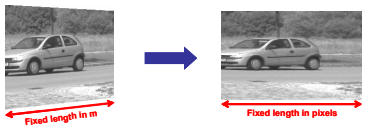

Fig. 8. Perspective reconstruction of the object side: Deformations as well as size differences caused by different distances to the video sensor are compensated.

In general, real objects do not exactly fit the cubical object model. This results in errors in the appearance of the reconstructed object sides. If the reconstructed object is not a vehicle, it can appear in any three dimensional shape. Therefore, the reconstruction error can be much higher than for cars. Fortunately, it does not seem to be likely that the reconstructed side of such an object is similar to a side view of a car.

In the case of vehicles, the error of the reconstruction is quite small. Thus, the reconstructed side is quite similar to the real object side. All background pixels and object parts, which are not aligned to the cuboid's side, are deformed in an undefined way. Strange deformations are possible as shown in the left part of the warped image. In addition, there even can be some errors at the reconstructed object side. This is due to the fact that car sides are not completely vertical as assumed by the object box model. Especially the front of a car often can only be reconstructed below the hood. However, the perspective warping creates object side views, which are much more similar to the original view than the mixture of deformed views in the video image.

The reconstructed views can now be used for the cascaded pattern recognition algorithm described above, which detects objects based on shape and contour information. Two cascades are applied. The first one detects car front and rear views and the second one side views.

D. Feature Extraction

The object detection with two cascades already performs excellent vehicle detection. Unfortunately, it is subject to several restrictions. The video sensor's field of view is rather small and the detection system can not classify objects, which are occluded or outside the ROI. In addition, the detector only performs the object detection for one object class. In order to create a consistent vehicle environment representation with multiple classes, a fusion of the single detector's results with the laser scanner features seems to be necessary.

For this purpose, several features in terms of floating point values are calculated. The primary feature describes the current output of one cascade. All object hypotheses in the evaluated ROI of an object are considered. The maximum stage s_{max} of the cascade, which is passed by at least one hypothesis, defines the main rating of this feature. If the last stage s_{last} is passed, the feature value will be 10. If only the stage before the last is passed, the feature value will be 9 and so on. The minimum feature value is 0.

An offset to the primary feature is calculated by considering multiple detections, which are usually generated by the cascaded object detector in cases of positive objects. The number of hypotheses h, which pass the maximum stage, defines an offset between 0 and 1. Due to experimental results, a maximum of 20 passing hypotheses is considered. Each hypothesis, which passes the maximum stage, increases the feature value by 0.05.

The detector can not be applied, if the object is occluded or not in the field of view. In this case, the feature value will be -1. The primary feature f is calculated by:

$$f = \begin{cases} -1 & \text{, if occluded or not in the FOV} \\ \max(0, s_{last} - s_{\max} + \min(1, 0.05 \cdot h)) & \text{, if detector was applied} \end{cases}$$

The described feature type is calculated for each detector. The maximum of the two feature values of both vehicle detectors defines a third feature.

Since the objects are often occluded in the field of view of the video sensor, the temporal maxima of the three features are stored as three additional features.

VI. Classification

The feature level fusion is performed by the classification part of the framework. The layout of this part is illustrated in Figure 9. A pattern classifier calculates membership values for each of the distinguished classes by applying statistical classification. The pattern classifier is based on neural networks. Details about this classifier were described in [13].

Afterwards, a rule based classification part is applied to verify and correct the output of the pattern classifier. The membership values are manipulated according to the rules. This part can also be used to guarantee a specific behavior of the system. A temporal mean filter stabilizes the output of the classification.

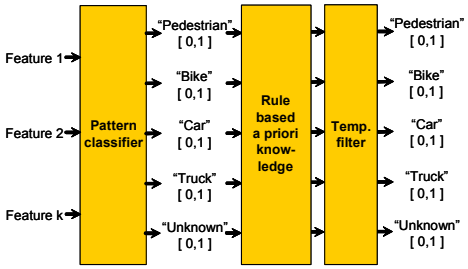

Fig. 9. Classification: All available features are combined by the pattern classifier, which calculates membership values for all distinguished classes based on statistical classification. The rule based part also evaluates the features and verifies and corrects the output of the pattern classifier. A temporal mean filter also includes classification results of former time steps.

VII. Output Adaptation

An advantage of the framework is the type of the classification output, which consists of the provided membership values. This output configuration allows for a simultaneous optimization of the membership values and the selected class depending on the different requirements of the applications [1].

VIII. Results

A. Evaluation Measures

The proposed system is evaluated with a set of labeled test data, which was not used for the training of the pattern classifier. The test set consists of several sequences of urban, suburban and highway scenarios with a total of approximately 50000 frames.

The application of classification is usually a tradeoff between low numbers of false alarms (false positives) and high numbers of correct detections (true positives). In statistics, the numbers of true and false positives are described with the following measures:

$$detection\ rate = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$false\ positive\ rate = \frac{false\ positives}{false\ positives + true\ negatives}$$

Different combinations of both measures can be plotted as a ROC (Receiver Operating Characteristic).

Unfortunately, the number of true negatives is not exactly defined for the vehicle environment perception task. The amount of background objects is usually much higher than the number of other classes' objects. Therefore, only objects of the other classes are labelled. Furthermore, it is often not possible to estimate the correct number of background objects (i.e. when the sensors measure areas with bushes, woods or gardens). For this reason, another measure is used to describe the number of false alarms:

$$false\ detection\ rate = \frac{false\ positives}{false\ positives + true\ positives}$$

This measure describes the ratio of false alarms to all detections. Operating Point Curves (OPC) will show the detection rate over the false detection rate.

B. Vehicle Detector

The first evaluation only evaluates the performance of the car detection system, which consists of two cascaded classifiers. The corresponding operating point curve for the test data is shown in Figure 10.

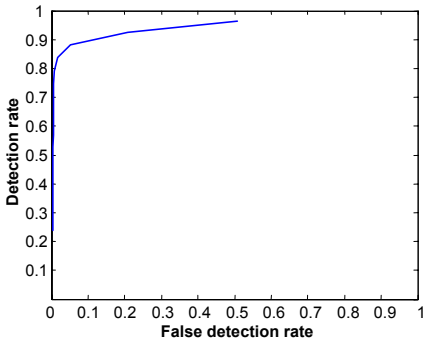

Fig. 10. Operating point curve for the video car detector: This plot shows the detection rate over the false detection rate.

The video car detector already performs very well, but due to the small field of view of the camera it is not possible to detect all cars in the test vehicle's environment. Therefore, the feature level fusion system is designed to present a more complete vehicle environment model.

C. Complete Fusion System

The fusion system was also evaluated with the test data in the complete field of view of the laser scanner. Figure 11 shows two Operating Point Curves. The blue dotted curve shows the performance of the system, which only uses a laser scanner. The red curve shows the curve of the complete feature level fusion system. Obviously, the pure laser scanner based classification already performs very well. The improvement of the feature level fusion is rather small. There are several reasons for this small improvement.

Firstly, the field of view of the video camera is much smaller than the laser scanner's field of view. Thus, the classification can only be improved for some of the objects. Secondly, the laser scanner already performs a good classification of moving objects. Detailed analyses of the test sequences have shown that the achieved benefit is primarily on non-moving cars, but the evaluated scenarios contain a much higher amount of moving cars than non-moving cars.

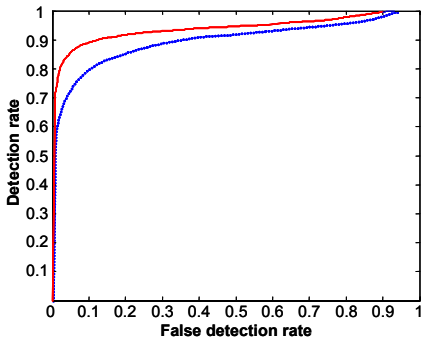

Fig. 11. Operating point curve for the complete system: The blue dotted line shows the performance of the pure laser scanner based approach. The red line shows the performance of the feature level fusion approach.

However, the small amount of cases, which are improved by the video camera, is nevertheless of importance. Non-moving objects, which are poorly classified by the pure laser scanner approach, can especially cause dangerous situations, if they are located in front of the test vehicle and therefore in the field of view of the video camera.

D. Processing Time

The complete system was applied in real time at the sensor measurement frequency of 12.5 Hz on an Intel Pentium 4 3.2 GHz desktop computer.

IX. Conclusion

A real time feature level fusion system for object classification was introduced. The approach was implemented and analyzed. While the extracted video features at the moment were only evaluated for the classification of cars, other works already have demonstrated that the described video feature extraction will also work for other classes. First experiments with a cascaded pedestrian detector already showed promising results and have to be evaluated in future works.

The statistical improvement of the fusion approach compared to the pure laser scanner based approach is rather small, but detailed analyses of the improved cases have shown, that the video sensor can improve the classification of non-moving objects, which can be of outstanding importance, as they are directly located in front of the test vehicle.

References

[1] S. Wender, K. C. J. Dietmayer: "An Adaptable Object Classification Framework", in *Proceedings of 2006 IEEE Intelligent Vehicles Symposium*, Tokyo, Japan, 2006

[2] N. Kaempchen, K. C. J. Dietmayer: "Fusion of Laserscanner and Video for Advanced Driver Assistance Systems", *in Proceedings of 11th World Congress on Intelligent Transportation Systems*, Nagoya, Japan, 2004

[3] N. Kaempchen, M. Buehler, K. C. J. Dietmayer: "Feature-Level Fusion for Free-Form Object Tracking using Laserscanner and Video", in *Proceedings of 2005 IEEE Intelligent Vehicles Symposium*, Las Vegas, USA, 2005

[4] S. Wender, T. Weiss, K. Fuerstenberg, K. C. J. Dietmayer: "Object Classification exploiting High Level Maps of Intersections", in *Proceedings of Advanced Microsystems for Automotive Applications 2006*, Berlin, Germany, 2006

[5] S. Wender, M. Schoenherr, N. Kaempchen, K. C. J. Dietmayer: "Classification of Laserscanner Measurements at Intersection Scenarios with Automatic Parameter Optimization", in *Proceedings of 2005 IEEE Intelligent Vehicles Symposium*, Las Vegas, USA, 2005

[6] P. Viola, M. Jones: "Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade", *Advances in Neural Information Processing Systems 14*, MIT Press, 2002

[7] I. Kallenbach, R. Schweiger, G. Palm, O. Loehlein: "Multi-class Object Detection in Vision Systems Using a Hierarchy of Cascaded Classifiers", in *Proceedings of 2006 IEEE Intelligent Vehicles Symposium*, Tokyo, Japan, 2006

[8] P. Viola, M. J. Jones, D. Snow: "Detecting Pedestrians Using Patterns of Motion and Appearance", in *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003)*, Nice, France, 2003

[9] S. Wender, O. Loehlein: "A Cascade Detector Approach Applied to Vehicle Occupant Monitoring with an Omni-Directional Camera", in *Proceedings of 2004 IEEE Intelligent Vehicles Symposium*, Parma, Italy, 2004

[10] M. Maehlisch, R. Schweiger, W. Ritter, K. Dietmayer: "Sensorfusion Using Spatio-Temporal Aligned Video and Lidar for Improved Vehicle Detection", in *Proceedings of 2006 IEEE Intelligent Vehicles Symposium*, Tokyo, Japan, 2006

[11] D. Ponza, A. Lopez, F. Lumbreras, J. Serrat, T. Graf: "3D Vehicle Sensor based on Monocular Vision", in *Proceedings of IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria, 2005

[12] H. Schneiderman, T. Kanade: "A Statistical model for 3D Object Detection Applied to Faces and Cars", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2000

[13] S. Wender, K. C. J. Dietmayer, "Statistical Approaches for Vehicle Environment Classification at Intersections with a Laserscanner", in *Proceedings of ITS 2005, 12th World Congress on Intelligent Transportation Systems*, San Francisco, USA, 2005