Multimodal Highway Monitoring for Robust Incident Detection

Michael Pucher, Dietmar Schabus, Peter Schallauer, Yuriy Lypetskyy,
Franz Graf, Harald Rainer, Michael Stadtschnitzer, Sabine Sternig,
Josef Birchbauer, Wolfgang Schneider, Bernhard Schalko

***Abstract*— We present detection and tracking methods for highway monitoring based on video and audio sensors, and the combination of these two modalities. We evaluate the performance of the different systems on realistic data sets that have been recorded on Austrian highways. It is shown that we can achieve a very good performance for video-based incident detection of wrong-way drivers, still standing vehicles, and traffic jams. Algorithms for simultaneous vehicle and driving direction detection using microphone arrays were evaluated and also showed a good performance on these tasks. Robust tracking in case of difficult weather conditions is achieved through multimodal sensor fusion of video and audio sensors.**

I. INTRODUCTION

To control the increasing traffic-flow on highways and to meet safety and security standards, monitoring of traffic is becoming more and more important. For this purpose the Austrian road operator ASFINAG has around 600 cameras on high- and expressways in open road surroundings. To improve the workflow of the operators and to ensure that almost every accident is recognized, video-based incident detection is required. Existing incident detection systems have to be improved for the reliable surveillance of highways. Especially wrong-way drivers must be detected reliably to avoid severe damages to health and property.

Traffic safety can be improved through the increase of detection rates and the decrease of false alarm rates of automatic event detection systems. These systems form the basis for a wide number of possible applications. In our work we focused on three scenarios, namely detection of wrong-way drivers, traffic jams, and still standing vehicles.

Operators watching video surveillance cameras are the simplest and most common method to monitor roads and highways. Detecting wrong-way drivers using this approach is difficult, since the personnel have to watch the cameras all the time to spot the event. For such critical events like wrong-way drivers, feedback has to be given within a very short time interval to be able to act properly.

Automatic traffic surveillance methods are therefore an appropriate method to overcome this drawback, since information is retrieved instantly or with a sufficiently small latency. Today many vehicle detection systems rely on inductive loop detectors and TriTech (infrared, ultrasonic, radar) sensors. However, installation and maintenance problems of these detectors have necessitated the development of non-intrusive alternative solutions with low maintenance costs. A few non-intrusive systems have become more prominent in the last years on Austrian highways, where in most cases Tri Tech sensor techniques were applied.

Starting from the existing methods and systems, we improved and adapted these systems in order to get more robust systems. The major focus of our work was to improve the video system and to increase the robustness of the whole system in case of fog, occlusion, or poor sight using audio-based methods and multimodal sensor fusion.

Audio-visual detection methods have the advantage of being non-intrusive, i.e., no sensors have to be installed on the road surface. Furthermore, video cameras are already ubiquitous such that a video-based detection system can cover a wide area of highways without the need to install additional sensors.

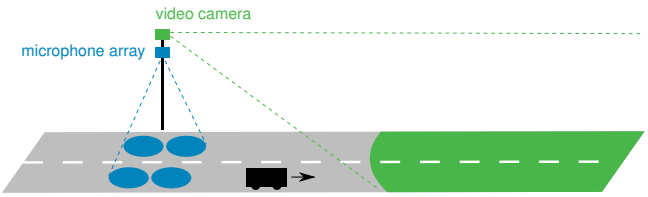

Fig. 1. Scenario for audio-visual highway monitoring

Fig. 1 shows the basic system setup of our test site. It is equipped with video cameras and acoustic sensors (microphone arrays), which are mounted on a highway gantry. The video cameras are able to monitor a relatively large area of the highway and are therefore used for the detection of wrong-way drivers, still standing vehicles, and traffic jams. The microphone array is able to monitor a relatively small area of the highway and is therefore used for the detection of wrong-way drivers only, as the other two events cannot reliably be detected in such a small area. To fuse the two sensors, we have to apply temporal fusion since the areas

M. Pucher and D. Schabus are with Telecommunications Research Center Vienna, Donau-City-Straße 1, 1220 Vienna, Austria {pucher, schabus}@ftw.at.

P. Schallauer, Y. Lypetskyy, F. Graf, H. Rainer, and M. Stadtschnitzer are with Joanneum Research Forschungsgesellschaft mbH, Steyrergasse 17, 8010 Graz, Austria {peter.schallauer, yuriy.lypetskyy, franz.graf, harald.rainer, michael.stadtschnitzer}@joanneum.at.

S. Sternig is with Graz University of Technology, Institute of Computer Graphics and Vision, Inffeldgasse 16, 8010 Graz, Austria sternig@icg.tugraz.at.

J. Birchbauer und W. Schneider are with Siemens AG Österreich, Corporate Technology (CT T CEE), Strassganger Strasse 315, 8054 Graz, Austria {josef-alois.birchbauer, wolfgang.b.schneider}@siemens.com

B. Schalko is with ASFINAG Maut Service GesmbH, Am Europlatz 1, 1120 Vienna bernhard.schalko@asfinag.at

monitored by the two sensors are not the same and do not even overlap. Sensor fusion is used to realize a more robust system that can maintain the information of the video system and thereby monitor a large area of the highway. To realize temporal fusion it is necessary to relate the two different detection regions and make assumptions about the behavior of vehicles in the area which is not monitored. In our case, we use the velocity estimation from the video system and assume constant velocity of the vehicles in the area between audio and video sensor.

II. VIDEO DETECTION

Within the video detection task we have to identify image regions that correspond to vehicles. Different strategies are required for the detection of cars and trucks. This is caused by different levels of complexity between the two vehicles types. Trucks have a greater intra-class variation and the features used for car detection cannot be transfered to represent these differences.

A. Car Detector

For car detection, a scene-specific real-time approach is used [1], [2]. This approach is based on classifier grids, where an input image is sampled with a fixed highly overlapping grid. Each grid element corresponds to one classifier. The idea of classifier grids is visualized in Fig. 2. The scene calibration, e.g., knowing the ground plane, is used to reduce the search space significantly and to ensure real-time performance.

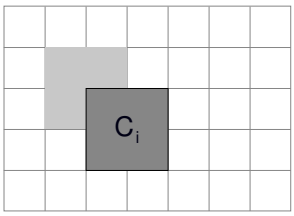

Fig. 2. The input image is divided into highly overlapping grid elements, where each grid element corresponds to one classifier.

As classifiers, on-line boosting for feature selection was used [3]. Boosting forms a strong classifier $H(x) = \sum_n^N \alpha_n h_n(x)$ through a linear combination of N weak classifiers $h_n(x)$ [3].

B. Truck Detector

Since more complex features are needed for truck detection, which cannot be used with classifier grids under real-time constraints, a different approach was used for truck detection. Here a background model was used to identify regions of interest, and within the regions of interest a support vector machine was used as a classifier in a sliding window manner.

An approximated median background model [12] was used, where the background template B for pixel (m, n) is updated according to

$$B_t(m,n) = B_{t-1}(m,n) + \text{sign}(B_{t-1}(m,n) - I_t(m,n)) * c.$$

The pixel intensity of the background model $B_t(m, n)$ is increased by c if the intensity of the current image $I_n(m, n)$ is less than the intensity in the background model, and decreased otherwise. Thereby regions can be detected that vary from the background model in order to identify large blobs with potential trucks. For the detection of trucks within these detected regions we use a support vector machine for classification [13].

III. VIDEO TRACKING

The vehicle detector described above serves as input for video based vehicle tracking. For performance reasons we apply the detector only in small detection areas (see Fig. 3) where vehicles enter the tracking area. Once a vehicle is detected in the detection area, it will be tracked independently from the detector. The vehicle tracking algorithm itself consist of three core parts.

The basic tracking approach allows efficient vehicle tracking near the camera. The feature points tracking algorithm described in [4] is used as input. In the whole image a predefined number of best trackable feature points is generated and tracked from frame to frame with sub-pixel accuracy. We exploit only feature points that could be tracked at least over the last N frames (typically, $N = 4$). This helps to reduce outliers due to noise or other video deteriorations. This feature point tracking is done at half image resolution by using every second field of the video. Our evaluations have proven that the tracking quality for tracking near the camera remains comparable with full resolution tracking, while the computing time is decreased significantly. At the same time we avoid the need for a de-interlacing scheme, which would be required for full resolution tracking of vehicles. Ground plane calibration is taken into account for calculating the vehicle region movement out of the motion of multiple feature points. The calibration will be done automatically once for each camera position.

As the second algorithmic core part, we propose the position and appearance check scheme. Since small frame-to-frame tracking errors could be accumulated over longer periods of time, potentially resulting in a position drift, we apply an appearance based algorithm partially described in [5]. We correct a tracked position with the help of an individual vehicle appearance model every N frames (typically, $N = 10 \ldots 15$). The algorithm is implemented with sub-pixel accuracy – this is especially important for areas distant from the camera where objects displacements between two successive frames can be significantly smaller than a pixel.

As a third algorithmic core part, we apply long distance tracking in areas far distant from the camera (see Fig. 3), where we use the full image size if the size of the tracked vehicle is becoming smaller than a certain threshold T ($T =$ 20 pixels). This allows us to increase the tracking distance by 140 meters on average as shown in Table I. Cars can be tracked at least 380 meters and up to 600 meters in maximum from the camera.

Fig. 3. Video-based vehicle tracking. Detection area indicated in white, tracking area in light grey, long-distance tracking area in black and the actual vehicle tracks are in dark grey.

IV. EVALUATION OF THE VIDEO SYSTEM

The video detection and tracking system was integrated into the Siveillance^{TM} hardware platform for testing and evaluation. Velocity estimation is also realized within this platform according to the method described in [7] based on pixel velocity and world coordinates. The platform was also used to evaluate the detection performance of the video system for the 3 types of events that are of interest to us, i.e., traffic jams, still standing vehicles, and wrong-way drivers. Precision and recall values for the different events are shown in Table II. Basis for the evaluation have been approximately 200 videos with an average length of 5 minutes, that have been acquired during the project period, mainly on the field test site location in Inzersdorf, Austria. The videos have been selected in such a way that they are challenging, e.g., by using an independent sensor system that is capable of detecting some of the events in the scope of our work.

Recall and precision values are defined as follows.

$$\text{Precision } [\%] = \frac{\#\text{Correct event detections}}{\#\text{Event detections in total}}$$

$$\text{Recall } [\%] = \frac{\#\text{Correct event detections}}{\#\text{Events in total}}$$

In words, precision gives the percentage of true events among all events reported by the system, and recall gives the percentage of true events that were correctly detected by the system among all true events. A perfect system would reach 100% for both values, but false alarms decrease precision and missed events decrease recall.

TABLE I

LONG-DISTANCE TRACKING.

Method	Car tracking distance in meters (min. / mean / max.)
No long-distance tracking	266 / 359 / 495
Long-distance tracking	380 / 499 / 602

TABLE II

VIDEO-BASED EVENT DETECTION.

Event	Recall / precision in %
Wrong-way driver	100 / 91
Still standing vehicle	93 / 95
Traffic jam	99 / 100

V. AUDIO DETECTION

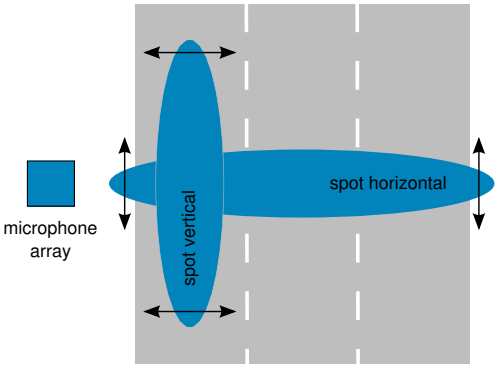

Fig. 4. Beam steering using horizontal and vertical line arrays.

Previous approaches to vehicle detection using microphone arrays were not able to detect vehicles and determine their direction on multiple lanes at the same time due to the assumption of a fixed analog time delay [6]. With today's signal processing capabilities it is however possible to steer the beam (spot) of the microphone array using variable digital time delays. This allows us to use a cross-shaped microphone array to perform horizontal and vertical beam steering as shown in Fig. 4.

The horizontal array was first fixed to the 0°-direction to form a narrow beam across the lane direction. If a car passes this spot, the power gathered by the beam increases significantly. By thresholding the power a car is detected. If a car is detected, the algorithm uses the information of the surrounding steered beam powers, by steering the array along the lane direction, of subsequent and earlier time frames to extract the cars direction. In a similar way, the vertical line array is steered across the lanes to extract the powers at the detection time. The lane is derived by extraction of the position where the power is forming a maximum.

The performance of this method on our recorded databases is shown in Table III. The performance of direction detection is directly transferable to the detection of wrong-way drivers. Direction detection was applied to all vehicles that were correctly detected by evaluating the temporal structure of the horizontal beam power using linear regression.

TABLE III

AUDIO-BASED VEHICLE AND DIRECTION DETECTION.

Dataset	Vehicle detection Rec. / Prec. in %	Direction detection Detection rate in %
1707 vehicles in 155 min.	96.4 / 98.6	99.7
830 vehicles in 70 min.	97.0 / 95.0	98.6

VI. SENSOR FUSION

Multisensor data fusion has been investigated for a long time [9], [14]. It remains an important topic since more different sensors and modalities are processed in today's information processing systems.

Sensor fusion of the visual and auditive modality has been investigated intensively within the speech processing community in audio-visual speech recognition [15] and speaker tracking [10], [16] but also in domain independent approaches to audio-video object localization [17]. Audio-visual sensor fusion for vehicle detection has been investigated in [18]. These previous works on audio-visual sensor fusion do however not consider temporal sensor fusion, where the observed areas between audio and video sensor do not overlap. This is the case for the highway monitoring scenario that we consider and also for many other different realistic scenarios. It is often not possible to mount video and audio sensors in a way such that temporal fusion can be avoided.

Temporal fusion of asynchronous sensors increases the complexity of the fusion system, since assumptions have to be made concerning the behavior of objects in areas which are not observed. [19] formulates a Bayesian approach to this problem for multiple sensors and targets in automotive applications.

We developed a sensor fusion system for the fusion of asynchronous sensors. For our first sensor fusion system, we decided to implement a *temporal decision level fusion* system with decision-in and decision-out [9].

By applying sensor fusion, we aim at combining the advantages of both sensors while getting rid of the disadvantages of the sensors. The principal advantages (+) and disadvantages (–) of both sensors, apart from practical advantages of the video sensor like availability of cameras at many places, are as follows:

- Video
  - \+ Observation of vehicles within a long time-space interval (hence suitable for all three scenarios)
  - – Performance depends on weather conditions
- Audio
  - – Observation of vehicles within a short time-space interval (hence suitable for the detection of wrong-way drivers only)
  - \+ Robust against different weather conditions
- Sensor fusion
  - \+ Observation of vehicles within a long time-space interval (hence suitable for all three scenarios)
  - \+ Robust against different weather conditions

Because we have a time span where the vehicle is not observed (see Fig. 1) we have to achieve a temporal sensor fusion. To correlate the video and audio detections / trackings we use the velocity estimated from the video tracking. Fig. 7 shows the correlation between estimated velocity on the x-axis and the number of frames between the appearance of the vehicle on the audio and video sensor on the y-axis. The

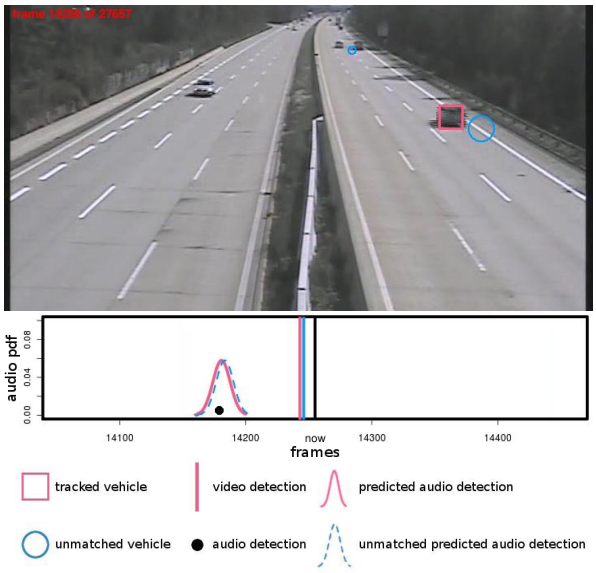

Fig. 5. Sensor fusion system with unmatched vehicle.

polynomial regression leaves a certain mismatch, which can have three reasons:

1) No constant velocity between video and audio event, contrary to assumption.
2) Error of video-based velocity estimation.
3) Differences of first appearance of video tracking. Video tracking may start on different frames depending where the first detections happens.

Fig. 6. Sensor fusion system with matched vehicles.

Although we have to deal with these sources of error, we are still able to match audio detections and video tracks as shown in Fig. 5 and Fig. 6. In Fig. 5 we have one audio detection, which is matched to one video track and one video track remains unmatched. In Fig. 6 both video tracks are matched to the corresponding audio detections. The sensor fusion algorithm performs an optimal matching of video tracks and audio detections using the predicted audio detections shown as bell curves at the bottom. It aligns events in a certain time interval by finding the best alignment.

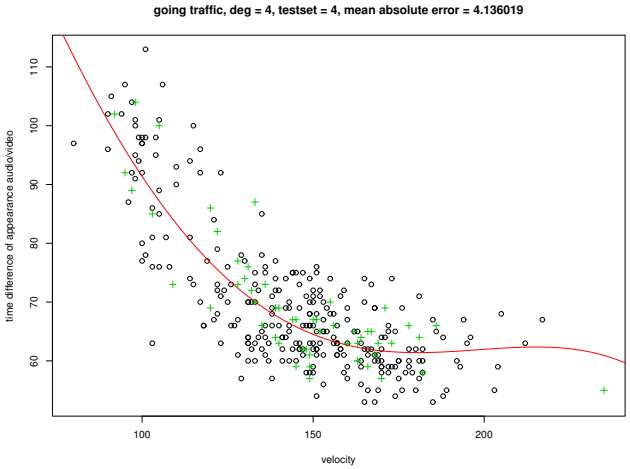

Fig. 7. Velocity to frame distance correlation.

Fig. 9. Sensor fusion system with no fog.

Fig. 8. Video system with simulated fog.

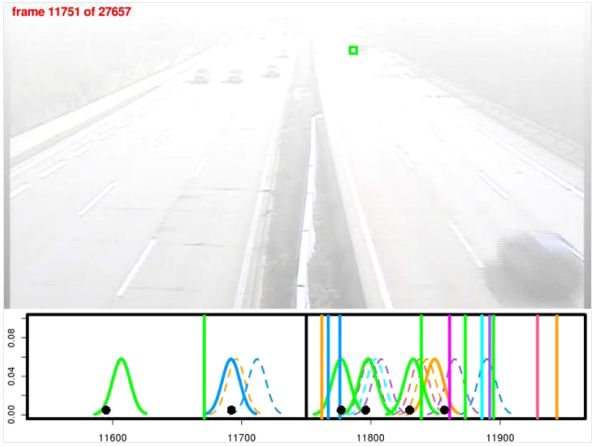

Fig. 10. Sensor fusion system with fog.

The biggest advantage of sensor fusion is realized under difficult weather conditions. Therefore we simulated fog by changing the contrast of videos continuously from the top to the bottom (see Fig. 8). This is a scenario which is very difficult for the video system, as can be seen by the low number of video tracks compared to audio detections in Fig. 8.

For the sensor fusion system we increased the sensitivity of the video system and applied temporal decision level fusion using this modified video system. Using the more sensitive system, we get a higher number of video detections / trackings, which are matched to the audio detections as can be seen in Fig. 9 and Fig. 10. A higher number of detections is necessary to be able to improve the baseline video system. Otherwise it would not be possible that the sensor fusion system outperforms the video system when fog is present.

The results for this system, the audio system, and the regular video system are shown in Table IV using cross validation on our data set. Note that the results in this table are not event detection results, but reflect the success of correctly detecting/tracking vehicles. Even though the results for the audio system alone are better than for the sensor fusion system, the audio system is by itself of no use to detect, e.g., still-standing vehicles. Only through successful combination with a track from the video system can we hope to detect such an event. In comparing the sensor fusion system with the video system, we can see that the former is more robust in the presence of fog, where the latter fails: only 12% of all passing vehicles are detected. By combining the two sensor modalities, we bring the robustness of the audio system to all three scenarios, reaching 89% of recall at the vehicle detection level under the challenging fog condition.

TABLE IV
SENSOR FUSION BASED TRACKING (320 VEHICLES IN 20 MIN).

Method	No fog Rec. / Prec. in %	Fog Rec. / Prec. in %
Audio	93 / 92	93 / 92
Video	100 / 47	12 / 95
Sensor fusion	86 / 86	89 / 90

In this system design, the audio detection is still a baseline for the sensor fusion system, which cannot be surpassed since we perform decision level fusion. In the future we plan to implement score level fusion to overcome this constraint.

VII. CONCLUSIONS AND FUTURE WORK

A. Conclusions

We have shown how to implement a robust real-time video based event detection system and have evaluated the system on a realistic data set taken from recordings that were made on Austrian highways. The system was evaluated on the detection of three critical traffic events, namely *traffic jam*, *still standing vehicle*, and *wrong-way driver*. The system was shown to have very good performance on the detection of these events. We evaluated long-distance tracking and showed that we can track vehicles between 200 and 600 meters and also realized video-based velocity estimation.

We have developed methods for the acoustic detection of wrong-way drivers. These methods were also evaluated on realistic audio recordings from highways, where they showed good performance on vehicle and direction detection. Especially for bad weather conditions the audio system is very valuable for detecting wrong-way drivers.

We have shown how sensor fusion can be used to make the system robust in the case of difficult weather conditions, while preserving the full information from the video tracking. The sensor fusion algorithms were also evaluated on realistic synchronized audio-video recordings from highways. The sensor fusion system which was described in this paper is a first step towards realizing the full potential of sensor fusion for highway traffic monitoring.

All information that can be derived automatically from video, audio, and other sensor data can be used by the operator to inform drivers on the highways using the public highway information system. Thereby traffic safety and security of highways can be improved.

B. Future Work

In our future work we want to concentrate on different weather conditions (fog, rain, snow, etc.) and also include new scenarios (lost payload, persons walking on the highway, etc.). Furthermore, we also want to focus on difficult lighting conditions such as are present at night when there is no road lighting or when there are vehicles coming towards the camera. The tracking algorithms will also provide the basis for visibility estimation. Generally we aim at improving the robustness of our tracking and detection methods.

Concerning audio detection, we aim at realizing real-time algorithms that can be deployed on a highway network. To achieve this goal we will also investigate the design and development of low-cost microphones. Additional goals are the evaluation of audio detection algorithms under difficult weather conditions (rain, snow) and the optimization of vehicle and direction detection.

For sensor fusion, we will include information on the score level from video, audio, and additional sensors (Bluetooth; Combination of passive infrared, ultrasonic and microwave radar). Furthermore, we want to investigate the fusion of sensors with the same modality, e.g., video cameras directed towards different spots of the highway. Another interesting line of research is the integration of signaling information from a mobile phone network. This type of signaling data shall be fused with other sensor data for deriving high-level traffic events.

VIII. ACKNOWLEDGMENTS

This work was supported by the Austrian Government and the City of Vienna within the competence center program COMET and by the companies Asfinag Mautservice GmbH and Siemens Österreich AG within the COMET project *Highway Monitoring (HI-MONI)* [11].

REFERENCES

[1] S. Sternig, P.Roth, H. Grabner, and H. Bischof, "Robust Adaptive Classifier Grids for Object Detection from Static Cameras", *in Proceedings Computer Vision Winter Workshop 2009*, 2009.

[2] P. Roth, S. Sternig, H. Grabner, and H. Bischof, "Classifier Grids for Learning Robust Adaptive Object Detectors", *In Proceedings CVPR 2009*, 2009.

[3] H. Grabner and H. Bischof, "On-line boosting and vision", Proc. of CVPR 2006, volume I, pages 260-267. IEEE CS, 2006

[4] J. Shi and C. Tomasi, "Good Features to Track", IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, 1994.

[5] J. H. Lee and J. B. Ra, "Block motion estimation based on selective integral projections", *Proc. IEEE Internat. Conf. Image Process.*, Rochester, NY, pp. 689-692, Sept. 2002.

[6] E.F. Berliner, "Acoustic Highway Monitoring", US Patent Nr. 6,195,608, February 2001.

[7] A. Criminisi, I. D. Reid, and A. Zissermann, A plane measuring device, *Image Vision Computing*, 17(8), 625-634, 1999.

[8] N. Strobel, S. Spors, and R. Rabenstein, Joint Audio-Video Object Localization and Tracking, *IEEE Signal Processing Magazine*, January 2001.

[9] P. K. Varshney, Multisensor Data Fusion, *Electronics and Communications Engineering Journal*, vol. 9, pp. 245-253, December 1997.

[10] M.J. Beal, H. Attias, and N. Jojic, "Audio-Video Sensor Fusion with Probabilistic Graphical Models", *Proc. ECCV*, 2002.

[11] Highway Monitoring (HI-MONI), http://hi-moni.ftw.at.

[12] N. J. B. McFarlane and C. P. Schofield, "Segmentation and tracking of piglets", *Machine Vision and Applications*, 1995.

[13] V. N. Vapnik, "The Nature of Statistical Learning Theory", *Springer*, 1995.

[14] D.L. Hall, An Introduction to Multisensor Data Fusion, *Proc. of the IEEE*, 85(1), January 1997.

[15] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, Audio-visual speech recognition, Technical Report, John Hopkins Workshop, October 2000.

[16] M. Kpesi, M. Neffe, T. Van Pham, M. Grabner, H. Grabner, and A. Juffinger, "Audio-Visual Feature Extraction for Semi-Automatic Annotation of Meetings", *Proc. of MMSP'06*, Victoria, BC, Canada, October 2006.

[17] N. Strobel, S. Spors, and R. Rabenstein, Joint Audio-Video Object Localization and Tracking, *IEEE Signal Processing Magazine*, January 2001.

[18] A. Klausner, A. Tengg, and B. Rinner, "Vehicle Classification on Multi-Sensor Smart Cameras using Feature and Decision Fusion", Proc. of the International Conference on Distributed Smart Cameras (ICDSC'07), Vienna, Austria, September 2007.

[19] C. Coue, T. Fraichard, P. Bessiere, and E. Mazer, "Using Baysian Programming for Multi-Sensor Multi-Target Tracking in Automotive Applications", *Proc. of Int. COnf. on Robotics and Automation*, Taipeh, Taiwan, May 2003.