Revised Version (ISSI-2009-09-0154)

Autonomous Audio-Supported Learning of Visual Classifiers for Traffic Monitoring

Horst Bischof[1], Martin Godec[1], Christian Leistner[1],
Bernhard Rinner[2] and Andreas Starzacher[2*]

[1]Graz University of Technology, AUSTRIA
[2]Klagenfurt University, AUSTRIA

January 2, 2010

Abstract

Automated traffic monitoring plays an important role for increasing safety and throughput. However, most of the currently deployed systems only capture data from the traffic sensors, and human supervision is required for the traffic assessment. For automatic classification, there exist powerful visual and acoustic learning approaches, but these algorithms require a huge amount of hand-labeled data to obtain high accuracy.

In this paper we focus on autonomous visual detection and classification of vehicles. We propose a self-learning framework with the goal of significantly reducing the effort for manual configuration which is important for mobile and flexible platforms. Our system consists of a robust on-line boosting classifier that allows for continuous learning and concept drift. The learner is also less susceptible to class-label noise which is hard to avoid in real-world self-learning applications. Furthermore, we incorporate an audio sensor as an additional complementary source into the training process. This audio sensor source acts as teacher for the self-learning of the primary visual classifier and helps to resolve ambiguities typically present in single sensor settings. We implemented our framework on an embedded platform to support mobile and autonomous traffic monitoring and show that our approach is able to yield high performing visual vehicle detectors without hand-labeling any video data.

Keywords: vehicle classification; on-line learning; autonomous traffic monitoring; audio and video processing

*Authors in alphabetical order.

1 Introduction

The steady increase of automobiles in operation impacts our life in several ways. Road congestions induce severe economic consequences due to delays and energy waste; estimations on the total cost of congestions range up to 1% of the GDP. According to the European Transport Safety Council[1] some 39.000 people were killed in road collisions in 2008 in Europe. Hence, increased safety and throughput on the existing road infrastructure is a major concern.

Automated traffic monitoring plays an important role for increasing safety and throughput. Numerous sensors along the roads capture traffic data which is analyzed in order to assess the current situation. This assessment can then trigger various counter actions such as warning drivers, reducing speed limits or re-routing traffic. Given the huge scale and complexity of the traffic system we would like to automate traffic monitoring and control as much as possible. However, most of the current traffic monitoring systems only capture data from the traffic sensors; continuous human supervision is required for the assessment (cp. Sidebar on Intelligent Traffic Monitoring). Additionally, there is an increasing demand for mobile or portable monitoring systems needed to monitor temporary events such as construction sites.

Meanwhile, there exist both powerful visual and acoustic classifiers. However, in order to obtain high accuracy these algorithms require a huge amount of hand labeled data. Collecting this data is a tedious and cost-intensive task. The classifiers are usually trained in the lab and are later applied (without adaptation) to a wide variety of possible scenarios and might thus become unnecessarily complex. Additionally, typical appearance-based classifiers [1] are sensitive to the orientation of the object, which makes it also difficult to obtain well-performing general detectors. In contrast, specialized detectors for specific scenes promise to perform better in terms of both accuracy and efficiency. Since the complexity of the task is reduced for specialized detectors, the required amount of labeled training samples can be drastically reduced as well.

For practical application, these specialized detectors have to fulfill several requirements: First, they have to be able to train as autonomous as possible in order to avoid the human labeling effort for every site. Second, this autonomous learning has to be performed continuously in order to allow for varying scenario conditions such as weather and illumination changes. Finally, these systems have to be resource-effective in order to enable wide-spread usage.

In this paper we focus on autonomous visual detection and classification of vehicles. Several traffic monitoring systems exploit data from multiple and/or heterogeneos sensors (e.g., [2, 3, 4]). In contrast to these approaches, we propose a self-learning framework with the goal of significantly reducing the effort for manual configuration. In practice, however, the incorporation of noisy data can be hardly avoided for autonomous self-labeling systems, i.e., identifying false class labels. If these false labels accumulate over time in the learning process they can easily lead to drifting.

Our system achieves robustness through first applying a robust on-line boosting classifier that also allows for continuous learning in order to train a visual appearance-based detector. Second, we incorporate an additional complementary sensor source (i.e., audio classification) into the learning process. The audio classifier acts as an autonomous supervisor. It is initially trained on a small set of labeled data and supports the visual on-line classifier in its continuous self-learning process. The audio classifier achieves sufficient accuracy with only few training data and does not perform self-training which ensures stability. The audio classifier can also be interpreted as a generic prior applicable to many scenarios, which justifies one-time human labeling, while the visual detector is trained autonomously for each individual scene. Furthermore, we abstain from complex microphone arrays and calibrations, in practice usually necessary for audio classification. Our system uses a single consumer microphone acting as a teacher and complementary information source for the video classification in order to allow for reduced-costs, easy

[1] http://www.etsc.eu

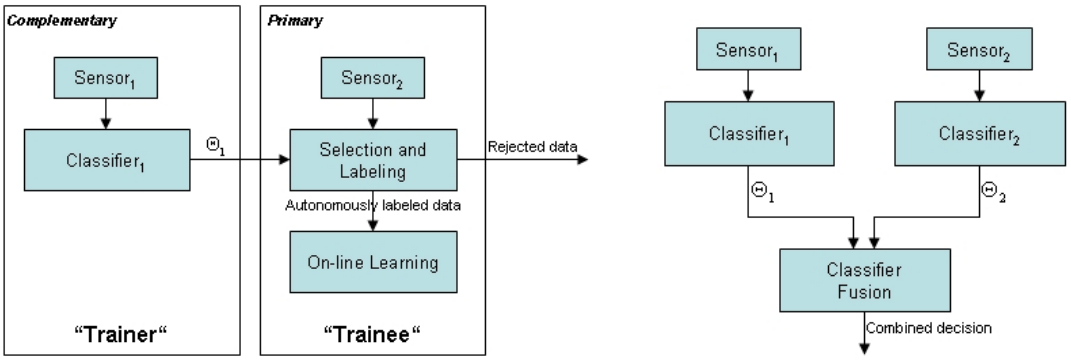


Figure 1: Self-training framework: the training process (left) and the classification process (right).

system deployment and maintenance. Another advantage of our approach is that we can use the audio classifier to resolve typical ambiguities between the vehicle classes, i.e., the classification between cars and trucks, which are hard to resolve for visual classifiers but are easy for acoustic classifiers.

In order to enable a mobile outdoor application, we have implemented our system on an embedded platform and demonstrate it for vehicle classification on highways using audio and image data. Our learning framework does not require any labeled visual data for on-line training and is able to improve the classification performance significantly.

2 Self-Training Framework

Fig. 1 depicts the overall structure of our autonomous self-learning framework. The left part describes the on-line training process using data from a primary and a complementary sensor source, respectively. The right part presents the collaborative classification process.

In the training process, both the audio and video sensors synchronously capture data of the observed scene. The complementary sensor acts as a trainer for the self-learning of the classifier of the primary sensor (trainee). The trainer's classifier is trained by using a small amount of hand-labeled audio data. For every detected object the trainer performs a classification—using the a priori trained classifier—and forwards a two dimensional parameter vector Θ_1 consisting of the decision (class label estimate) and its confidence value to the trainee. The trainee selects data only from objects with high classification confidence for its on-line training, i.e., it refuses objects and its associated data when the trainer's confidence value is below a threshold. For selected objects the trainee uses the trainer's classification result as label.

After the audio-supported on-line training of the visual classifier the trainer and trainee can be operated as independent classifiers. To improve the overall performance, we combine the output of both classifiers based on their confidences (cp. Fig. 1 right).

Note that our proposed system is similar to previous ones based on co-training [5], where two classifiers are first trained independently on labeled data and then train each other on unlabeled data. For instance, Levin et al. [6] trained a car detector using co-training [5] and Christoudias et al. [7] proposed an audio-visual co-training system for human gesture recognition. However, our system differs from both works in that (i) we use continuous on-line learning, (ii) we do not need any human labeling effort for the visual classifier and (iii) our audio classifier does never perform self-updates, which ensures long-term system stability. The latter argument is also supported by previous works which highlighted that co-training's main assumption, i.e.,

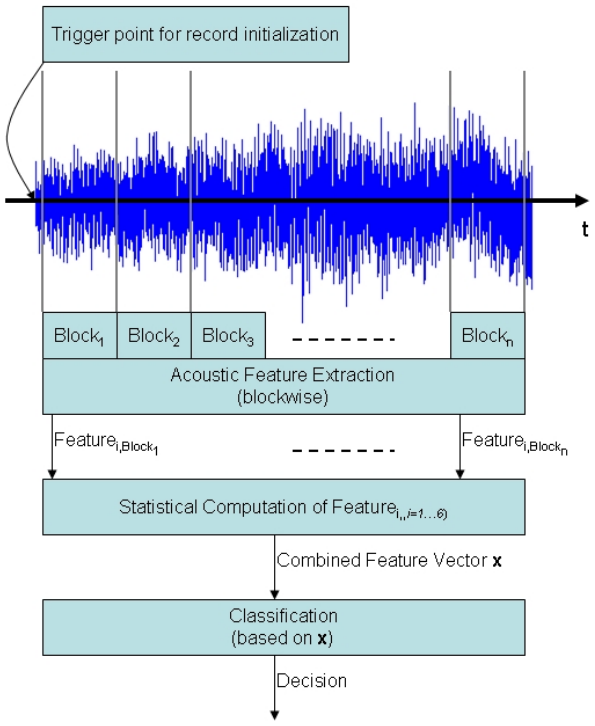


Figure 2: Acoustic classification system

conditional independence [5], is very hard to fulfill in practice and systems where an autonomous predictor in form of a classifier teaches another classifier have shown to perform better. For example, Roth et al. [8] used a generative model in order to conservatively update an on-line classifier and Wu et al. [9] trained an on-line classifier using an "oracle" for pedestrian detection. Our system differs from these two approaches in several aspects. First, we use an audio classifier as autonomous teacher. Second, we use robust on-line boosting as classifier and finally incorporate the teacher also in the final classification process in order to resolve ambiguities among vehicle classes.

3 Acoustic Classification

Fig. 2 depicts the basic structure of our acoustic classification system. A microphone captures the audio signal of passing vehicles along the road. In a first processing step we partition the audio samples into n blocks with a configurable block size. We then extract several acoustic features for each block individually. These block features are further abstracted into a single feature vector $\mathbf{x}$ by a statistical merging. The abstracted feature vector serves as input for the classifier.

The performance of the classification process strongly depends on the characteristics of the features. Our goal is to select a set of highly discriminative features for the considered classes. In our case we use a total of six different acoustic features which are defined in the time, spectral and cepstral domain, respectively [10].

The *short-time energy (ste)* is a simple time-domain feature which is highly discriminative between cars and trucks, but is also sensitive to noise. *Spectral bandwidth (spbw)*, *spectral roll-off point (spro)* and two coefficients of *band-energy ratio values* (ber_6 and ber_7) exploit different characteristics of the vehicle's emitted acoustic spectrum. The spectral bandwidth measures the

spread of frequencies around the spectral centroid. The spectral roll-off point indicates up to what frequency level a defined amount of percentage of the spectrum is accumulated. A higher roll-off value corresponds with more intense or higher frequencies. The band energy ratio values describe the ratio of energy in certain frequency subbands to the total signal energy. The ratio based on 6^{th} and 7^{th} subband yields more class-discriminative values than with the first five subbands. A *cepstral analysis (cep)* is performed as well.

The combined value of feature i is computed by statistically merging the $\text{Feature}_{i,Block_k}$ for all blocks. This task is performed for all different features described previously. Thus, the resulting features are combined into a six dimensional feature vector $\mathbf{x}$.

$$\mathbf{x} = (ste, spbw, spro, ber_6, ber_7, cep)^T \tag{1}$$

We implemented and evaluated several classification algorithms such as k-nearest neighbor (KNN), linear and quadratic discriminant analysis (LDA, QDA), naive Bayes (NBC), support vector machine (SVM) and artificial neural network (ANN) [11]. Each algorithm has its advantages and disadvantages depending on the dataset. Therefore, the choice of an algorithm is based on the specific application domain. The classification algorithms return the estimated class labels with their confidence values as output.

4 Video Classification

A common choice in visual traffic analysis is simple background modeling (BGM). However, a BGM has several disadvantages. For instance, it is sensitive to shadows, cannot discriminate between different vehicle classes and cannot detect vehicles in slow motion scenarios such as traffic jams.

For visual classification, we therefore train an appearance-based model avoiding the problems above. In particular, we follow the seminal work of Viola and Jones [1] who showed that cascades of boosted classifiers and efficient image representation (i.e., integral images) lead to real-time appearance-based object detection systems. However, our object detector differs in two aspects: First, we use on-line boosting for feature selection (i.e., [12]) to allow for continuous learning without storing any training samples. Second, we use more robust loss functions for on-line boosting which were recently proposed by Leistner et al. [13]. Using a robust learning algorithm is especially important in practice because label noise is an inherent problem in self-learning approaches.

In the training phase, we exploit the audio classifier (cp. Sec. 3) to extract training data from scene-specific video streams captured by a non-calibrated consumer camera. To avoid hand-labeling, we use a simple Gaussian background model [14] to extract initial motion blobs. Note that the BGM is just used to crop "regions of interest" for the training process of the boosting detector. During operation mode, only the appearance-based detector is used. To extract proper training blobs, we apply different kinds of post-processing such as size verification and positioning within the scene. Subsequently, we exploit the audio classifier which is able to separate these samples into scenarios containing single vehicles of either class and scenarios containing multiple vehicles or no vehicle at all. Note that we can also easily generate negative training examples from the scene with the audio classifier, i.e., we crop random patches from the scene if neither the BGM nor the audio classifier are indicating that there are vehicles. We train a car and a truck detector based on these patches.

Since most traffic applications are not only concerned in detecting vehicles but also in discriminating different vehicle classes, we train two different detectors—one for trucks and one for cars. To resolve visual ambiguities among the different vehicle classes, we also incorporate the acoustic classifier for making the final classification into either truck or car, since this is an

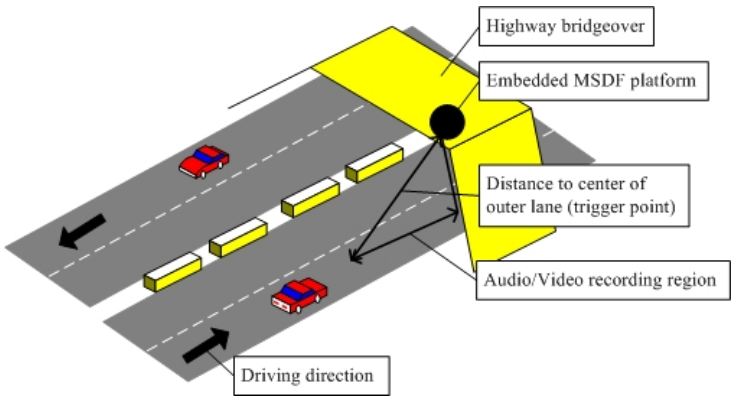


Figure 3: Experimental setup on a freeway with two lanes in both directions. Microphone and camera are connected to the embedded multi-sensor data fusion (MSDF) platform. The distance between the sensors and the outer lane is approximately 10 meters.

easier task for the audio classifier. We abstain from training a single detector for both cars and trucks because a high intra-class variance has to be covered, which usually leads to higher model complexity and thus slower detectors. Furthermore, we would loose the additional confidence provided by two visual detectors which can be coupled with the audio classifier.

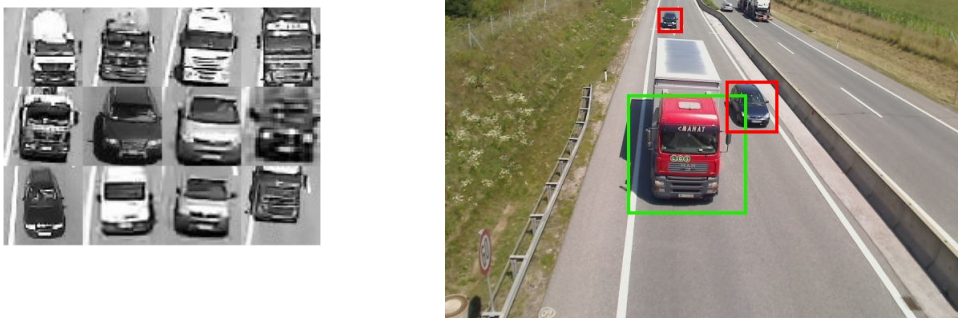
4.1 Collaborative Audio and Video Classification

During the classification phase, we unify the visual and the audio cue by linearly combining the confidences of both classifier types. To classify a scene, we first generate a visual classifier by applying our two visual detectors for cars and trucks to identify several candidate regions where at least one of the two detectors provides a positive confidence. Then, the confidences of the visual classifiers are combined with the confidences provided by the audio classifier (all confidences are normalized to the range of $[-1, +1]$ before fusion). In order to keep our approach simple, we use weighting parameters α and β for the combination of both confidences of the audio f_a and the visual f_v classifier. In particular, we use a simple arithmetic mean to weight the two confidences (i.e., both are set to $\frac{1}{2}$).[2] Finally, by using a non-maxima suppression the highest vote is estimated providing the according class for the candidate regions.

5 Experiments

Our experimental evaluation is based on real-world datasets of about 200 vehicles for each class (cars and trucks) from multi-lane freeway traffic. The datasets are partitioned into training and testing sets with 150 and 50 samples per class, respectively. Thus, we used 150 audio samples for each class to train the initial acoustic classifier. Fig. 3 depicts our experimental setup. The microphone was directed to the center of the outer lane. The audio data was recorded at 44.1 kHz in mono format with 16 bit resolution. The camera captured front shot images at a frame rate of about 5 Hz. Figure 4(a) shows some examples of cropped vehicle patches; Fig. 4(b) shows an example of the final detection and classification output. Video and audio recording were synchronized and started at a (virtual) trigger point. For each vehicle, sensor data of up to 4 seconds were captured—the actual recording period depended on the speed of the vehicles.

[2] α and β can be easily set to more "reasonable" values—for instance, by using cross correlation on labeled samples or using more sophisticated weighting techniques.



(a) (b)

Figure 4: Some automatically cropped sample patches (a) and final detection plus color encoded classification (b).

The experiments were performed on our MSEBX945 embedded computer board from Digital-Logic with a SMX945-L7400 CPU module. This platform provides interfaces to several sensory devices such as audio, video and laser sensors. The microphone is attached to a pre-amplifier from M-AUDIO which is connected to the embedded platform via USB. The camera is directly interfaced with the platform via FireWire over MiniPCI.

Our experimental evaluation aims at two goals. First, we want to show that our autonomous framework enables on-line training of classifiers under real-world conditions without any hand-labeling of the visual data. Second, we want to demonstrate that a collaborative classification of multiple sensors can gain significant performance improvements. For self-learning we use the audio sensor as trainer; for classification we exploit both cues.

In previous work [10] we showed that acoustic classifiers based on the feature vector given in Eq. 1 achieve notable classification accuracies of up to 93.75% with quadratic discriminant analysis (QDA). The other algorithms mentioned in Section 3 achieved about 90% (for ANN, SVM and LDA), 86.25% (for KNN) and 85% (for NBC). All of these results were obtained by 5-fold cross-validation with the datasets mentioned previously. Thus, we use the QDA classifier as trainer for our learning framework.

5.1 Autonomous Learning of Visual Classifiers

In our first experiment we trained two vehicle detectors—one on car and the other on truck samples, respectively. For representation, we use simple Haar-like features similar to [1] but abstain from training cascades, because the classifiers can be kept very simple due to their scene-specificity. For all experiments we used 100 selectors each with 50 weak classifiers. For the on-line boosting, we applied a logistic loss-function in form of $\log(1 + e^{-yF(x)})$ which has shown to be more robust than the exponential loss usually applied in on-line boosting [13]. We set the starting shrinkage factor s_{start} to 1, but decreased it with increasing number of selectors in the form of $s_t = \frac{s_{start}}{t+1}$.

As can be seen in Fig. 5(a), our system is able to train well-performing car and truck detectors without hand labeling of any visual data. To demonstrate the practical relevance of our approach, we performed a second set of experiments where we degraded the performance of our teachers (audio classifiers). In particular, we varied the noise level from 0% (perfect teacher without any misclassification) to 25% (teacher with 25% misclassification rate) which are ranges typically occuring in practice. As can be seen in Tab. 1(a) and Tab. 1(b), the recall rates hardly change with increasing noise level for both the car and the truck detectors, i.e., the

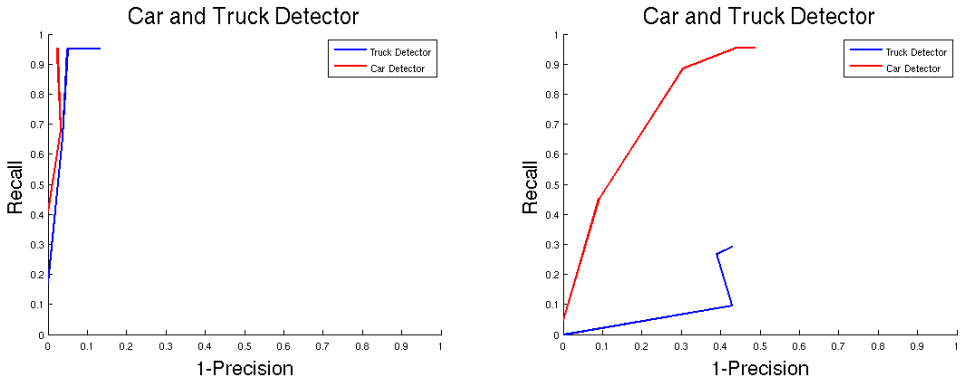
(a)

Noise	recall	precision	F-measure
0%	0.95 %	0.78 %	0.86 %
5%	0.95 %	0.48 %	0.64 %
10%	0.98 %	0.40 %	0.57 %
25%	0.98 %	0.34 %	0.50 %

(b)

Noise	recall	precision	F-measure
0%	0.98 %	0.17 %	0.29 %
5%	0.98 %	0.16 %	0.28 %
10%	1.00 %	0.15 %	0.27 %
25%	1.00 %	0.15 %	0.26 %

Table 1: Detector performance depending on different noise-levels for (a) cars and (b) trucks, respectively. As can be seen for cars, the recall rate stays very high even when the noise-level is at about 25%; however, the precision decreases. For trucks, increasing noise does not change the detection performance significantly. Please note that no postprocessing has been applied in this case. Classifiers have only been applied to the class they have been trained on.



(a) (b)

Figure 5: Automatically trained car and truck detectors. (a) The detectors achieve high classification performance when applied to test scenes only containing their training class. In this case, the detectors only discriminates the trained target class from the scene background. (b) When applied to scenes containing both vehicle classes, the performance degrades. The performance deteriorates dramatically, especially for the truck detections.

number of false-positives increases. The precision remains also constant for the truck detector, but the precision decreases with increasing noise for the car detector. However, we figured out in practice that if the recall rate stays high, a degraded precision can be corrected by applying smarter post-processing in case of multiple detections.

In the next two experiments we tested the car detector only on sequences with cars and the truck detector only on sequences with trucks, respectively (Fig. 5(a)). However, as can be seen in Fig. 5(b), the performance degrades dramatically if the two detectors have to cope with instances of both classes at the same time. Note that training the car detector using some truck samples as negatives and vice versa leads to a decreased recall while the precision can only be slightly increased. The main reason for this behavior is that especially the car detector is not able to discriminate parts of a truck to real cars, leading to a huge amount of false positives.

5.2 Collaborative Classification

In the third experiment, we used the same settings as above, but focused on a collaborative classification of audio and video. The idea is that the visual detector should be applied in order to locate the object in the video. Once an object has been detected, the audio classifier should support the visual detector in order to resolve ambiguities. In particular, after both

(a)

Classifier	recall	precision	F-measure
Truck	0.29 %	0.57 %	0.39 %
Car	0.95 %	0.51 %	0.67 %

(b)

Classifier	recall	precision	F-measure
Truck	0.85 %	0.71 %	0.78 %
Car	0.77 %	0.77 %	0.77 %

Table 2: Classification performance using (a) only visual classifier or (b) visual and audio classifier in combination. Comparing the F-Measure, which gives an impression of the overall performance, the improvement of the combined classification can be seen.

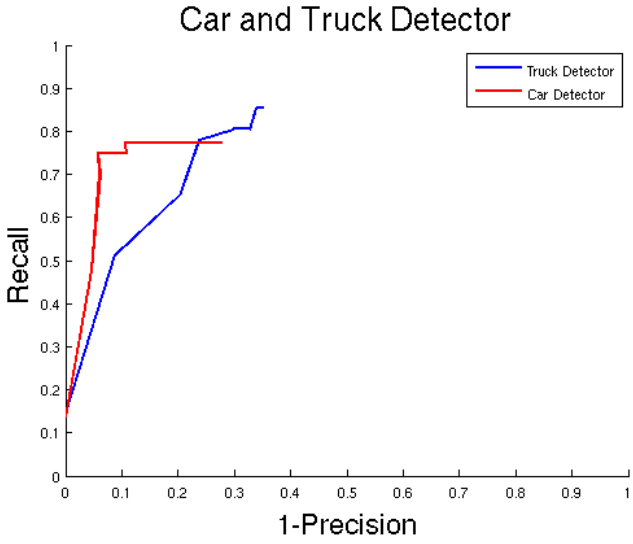


Figure 6: Final result of the collaborative classification using audio and video. The accuracy increases significantly if the audio classifier supports the visual detector in the final classification, especially for the truck detector.

visual detectors have been run over the video frame, we derive the final classification in a post-processing step by computing a linear combination of the video and audio classifiers as described in Section 4.1. Tab. 2 and Fig. 6 depict the result of this collaborative classification which leads to significantly improved detection results.

6 Conclusions

In this work, we have presented an on-line visual self-learning framework with audio support. An acoustic classifier using a single off-the-shelf consumer microphone acts as an additional complementary sensor source in the training process and reduces the sensitivity to noise of typical single sensor settings. Our approach does not need any calibration and can thus be applied in mobile, flexible, low-cost traffic surveillance platforms.

Although we have demonstrated our multi-sensor method for vehicle classification, self-learning is a general concept with high potential for many applications. We are confident that it may serve as an important step toward versatile, autonomous and intelligent traffic monitoring.

Acknowledgments

This work has been sponsored in part by the Austrian Research Promotion Agency under grant 813399.

References

[1] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, 2002.

[2] Andreas Klausner, Christian Leistner, Allan Tengg, and Bernhard Rinner, "An audio-visual sensor fusion approach for feature based vehicle identification," in *Proceedings of the 2007 IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2007)*, London, UK, Sept. 2007, pp. 1–6.

[3] V. Kastrinaki, M. Zervakis, and K. Kalaitzakis, "A survey of video processing techniques for traffic applications," in *Image and Vision Computing*, 2003, vol. 21, pp. 359–381.

[4] Alexandra Koutsia, Theodoros Semertzidis, Kosmas Dimitropoulos, Nikos Grammalidis, and Kyriakos Georgouleas, "Intelligent Traffic Monitoring and Surveillance with Multiple Cameras," in *Proceedings of the Sixth International Workshop on Content-Based Multimedia Indexing*, 2008, pp. 125–132.

[5] Avrim Blum and Tom Mitchell, "Combining Labeled and Unlabeled Data with Co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, Madison, WI, USA, 1998, pp. 92–100.

[6] A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using co-training," in *Proceedings of the Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003, vol. 2, pp. 626–633.

[7] C. M. Christoudias, R. Urtasun, A. Kapoor, and T. Darrell, "Co-training with noisy perceptual observations," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, FL, USA, 2009, pp. 1–8.

[8] P.M. Roth, H. Grabner, D. Skočaj, H. Bischof, and A. Leonardis, "Conservative visual learning for object detection with minimal hand labeling effort," in *Proceedings of the 27th DAGM Symposium*, Vienna, Austria, 2005, pp. 293–300.

[9] B. Wu and R. Nevatia, "Improving part based object detection by unsupervised, online boosting," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, 2007, pp. 1–8.

[10] Andreas Starzacher and Bernhard Rinner, "Single Sensor Acoustic Feature Extraction for Embedded Realtime Vehicle Classification," in *Proceedings of the 2nd International Workshop on Sensor Networks and Ambient Intelligence*, 2009, pp. 1–6.

[11] Andreas Starzacher and Bernhard Rinner, "Embedded Realtime Feature Fusion based on ANN, SVM and NBC," in *Proceedings of the 12th International Conference on Information Fusion (Fusion 2009)*, Seattle, USA, July 2009, pp. 1–8.

[12] H. Grabner and H. Bischof, "On-line boosting and vision," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, 2006, vol. 1, pp. 260–267.

[13] Christian Leistner, Amir Saffari, Peter M. Roth, and Horst Bischof, "On Robustness of Online Boosting - A Competitive Study," in *3rd IEEE On-line Learning for Computer Vision Workshop (ICCV'09)*, Kyoto, JP, 2009.

[14] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 1999)*, 1999, vol. 1, pp. 246–252.

Sidebar: Intelligent Traffic Monitoring

In the near future we will witness over a billion automobiles in operation worldwide [a]. Automated traffic monitoring will therefore play an essential role to improve the throughput and safety of roads. Current monitoring systems capture—usually vision-based—traffic data from a large sensory network; however, they require continuous human supervision which is extremely expensive. Future traffic monitoring systems must become more "intelligent" to analyze and assess traffic situations in real-time under virtually all weather conditions.

Robustness and adaptivity are key challenges for intelligent traffic monitoring. Numerous sensors are installed on various places, such as on poles, on gantries or even in the pavement, to capture the traffic and estimate different traffic parameters. This diverse setting typically requires tedious sensor calibration and adapting the analysis algorithms to the observed scenes. This calibration and adaptation should be done with as little human intervention as possible. On the other hand, robustness is a precondition for integrating traffic monitoring to various applications.

Research on intelligent traffic monitoring has been conducted for many years. Since it is widely recognized that image-based systems are flexible and versatile for advanced traffic monitoring applications, most research has focused on image and video analysis (e.g. [b, c, d]). Various image-analysis methods are applied to the data from individual cameras in order to estimate traffic parameters. These parameters can be related to individual vehicles such as detection, classification and tracking or to the traffic behavior over some period of time such as lane occupancy or travel time.

Another stream of research focused on improving the robustness by exploiting data from multiple sensors. Sensor fusion techniques are applied to exploit the different characteristics of homogeneous and/or heterogeneous sensors. Chellappa et al. [e] introduced a Markov Chain Monte Carlo technique for a joint audio/visual vehicle tracking. Acoustic beamforming estimates the direction of arrival which in turn guides the visual tracking. Klausner et al. [f] exploited acoustic and visual sensors for vehicle detection and classification by extracting discriminative features from the different sensors and performing sensor fusion at the feature- or decision-level, respectively. Kushwaha et al. [g] also exploited acoustic and visual information for vehicle tracking in urban environments. They perform multi-modal fusion on an embedded sensor network in an urban environment.

Recently several traffic monitoring systems have been deployed on a larger scale to evaluate automated traffic analysis under real-world conditions. Rodríguez et al. [h] describe a vision-based traffic monitoring system that is able to detect vehicles in real-time. The major objective is to tackle some of the challenges in real-world deployments such as shadows, occlusions, day and night transitions and slow traffic, that impede existing monitoring systems to achieve a stable accuracy in those situations. The proposed system works autonomously for a certain period of time without human intervention and has the ability to adapt automatically to several environmental conditions. Similarly, [i] proposes an example-based algorithm to detect moving vehicle in a vision-based traffic monitoring environment under changing conditions. The algorithm is designed to learn from examples. Hence, it does not need to incorporate any prior knowledge (prior vehicle model). The algorithm was evaluated under several varying environmental conditions and has achieved a satisfying performance. A real-time vision system for automatic traffic monitoring (VISATRAM) is presented in [j]. VISATRAM follows a 2D spatio-temporal image-based automatic traffic monitoring approach. The range of functions comprises vehicle counting, vehicle velocity estimation and classification using 3D measurements. Furthermore, Rigolli et al. [k] reinforce the need to improve road safety by investigating inferences about driver behavior and learning normal behavior driving modes. They propose an agent-based approach for analyzing the behavior of the drivers.

Sidebar References

[a] Hamid Gharavi, K. Venkatesh Prasad, and Petro Ioannou, "Advanced Automobile Technology (scanning the special issue)," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 328–333, February 2007.

[b] V. Kastrinaki, M. Zervakis, and K. Kalaitzakis, "A survey of video processing techniques for traffic applications," in *Image and Vision Computing*, 2003, vol. 21, pp. 359–381.

[c] Kai-Tai Song and Jen-Chau Tai, "Image-Based Traffic Monitoring With Shadow Suppression," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 413–424, 2007.

[d] Rita Cucchiara, Massimo Piccardi, and Paola Mello, "Image Analysis and Rule-Based Reasoning for a Traffic Monitoring System," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 2, pp. 119–130, 2000.

[e] Rama Chellappa, Gang Qian, and Qinfen Zheng, "Vehicle Detection and Tracking using Acoustic and Video Sensors," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, 2004, pp. 793–796.

[f] Andreas Klausner, Christian Leistner, Allan Tengg, and Bernhard Rinner, "An audio-visual sensor fusion approach for feature based vehicle identification," in *Proceedings of the 2007 IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS 2007)*, London, UK, Sept. 2007, pp. 6.

[g] Manish Kushwaha, Songhwai Ohy, Isaac Amundson, Xenofon Koutsoukos, and Akos Ledeczi, "Target Tracking in Heterogeneous Sensor Networks Using Audio and Video Sensor Fusion," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Seoul, 2008, pp. 14–19.

[h] T. Rodríguez and N. García, "An adaptive, real-time, traffic monitoring system," *Machine Vision and Applications*, vol. 18, pp. 781–794, 2009.

[i] LinJ. Zhou, D. Gao, and D. Zhang, "Moving Vehicle Detection for Automatic Traffic Monitoring," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 1, pp. 51–59, 2007.

[j] Z. Zhu, G. Xu, B. Yang, D. Shi, and X. Lin, "VISATRAM: A real-time vision system for automatic traffic monitoring," *Image and Vision Computing*, vol. 18, pp. 781–794, 2000.

[k] M. Rigolli and M. Brady, "Towards a behavioural traffic monitoring system," in *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 2005, pp. 449–454.