Audio-Visual Co-Training for Vehicle Classification*

M. Godec and C. Leistner and H. Bischof
Graz University of Technology
Inffeldgasse 16, 8010 Graz, AUSTRIA
{godec,leistner,bischof}@icg.tugraz.at

A. Starzacher and B. Rinner
Klagenfurt University
Lakeside Park, 9020 Klagenfurt, AUSTRIA
{andreas.starzacher,bernhard.rinner}@uni-klu.ac.at

Abstract

In this paper, we introduce a fully autonomous vehicle classification system that continuously learns from large amounts of unlabeled data. For that purpose, we propose a novel on-line co-training method based on visual and acoustic information. Our system does not need complicated microphone arrays or video calibration and automatically adapts to specific traffic scenes. These specialized detectors are more accurate and more compact than general classifiers, which allows for light-weight usage in low-cost and portable embedded systems. Hence, we implemented our system on an off-the-shelf embedded platform. In the experimental part, we show that the proposed method is able to cover the desired task and outperforms single-cue systems. Furthermore, our co-training framework minimizes the labeling effort without degrading the overall system performance.

1. Introduction

Automated traffic monitoring plays an important role in increasing safety and throughput on the existing road infrastructure. Due to a steadily increasing number of such systems, human inspection of the acquired data will no longer be feasible in the near future. Nowadays there already exist powerful high-level traffic monitoring systems that allow for traffic jam prediction or toll collection. However, in order to deliver reliable predictions, these systems demand highly accurate vehicle detections and classifications. These classifiers should be based - at least partly - on video information, because this eases verification by human operators.

Training of visual object detectors is an active field of research and there already exist a huge variety of approaches. The most popular approaches use simple image filters such as Haar-like features [28] or histograms of oriented Gaussian [7] and then apply powerful machine learning techniques such as boosting or SVMs. Although these methods are able to deliver excellent results, they demand large amounts of usually hand-labeled data - a problem often neglected in the literature. Hand-labeling data is a tedious and cost-intensive task and handicaps the vast deployment and maintenance of modern detection systems. Furthermore, these detectors are usually trained in order to be applicable on a general class of scenes. In contrast, scene-specific classifiers have to solve an easier task and are thus more accurate and less complex which enables their applicability on low-cost embedded systems. Thus, the preferred system should be able to perform highly accurate, scene specific and adaptive classification by using only a minimal amount of labeled data for training.

To meet all these requirements, we propose an autonomous vehicle classification and detection system based on audio-visual co-training using low-cost consumer sensors that, additionally, avoids complicated calibration and expensive microphone arrays. Therefore, we train two heterogeneous classifiers on a small amount of labeled data and to co-train them on a continuous stream of unlabeled data in order to yield scene-specific and highly adaptive classifiers. Fusing audio and video information for traffic monitoring is not new and has been frequently proposed [15, 30, 14]. In these systems, typically information is combined at various levels of data abstraction such as raw data, features or decisions. The main objective is to exploit heterogeneous sensors to increase the robustness, confidence and the spatial as well as temporal coverage [25, 12]. Due to limited resource availability of our co-training framework a careful selection of appropriate algorithms has been performed.

Our system differs to most of the previous ones in terms that we do not use a fusion strategy on the decision level but to exploit the power of multiple sensors in order to perform robust autonomous learning. The most similar approach to ours is the recent work of Christoudias *et al.* [6] who performed audio-video co-training in order to learn hu-

*This work has been supported by the Austrian FFG projects MobiTrick (825840) and EVis (813399) under the FIT-IT programme and by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund, Interreg IV and the Carinthian Economic Promotion Fund KWF under grant KWF-20214/18354/27107.

man gesture recognition systems. However, [6] used off-line learning strategies which means that they exploit the entire training set at once which eases optimization and typically yields good results. In contrast, our approach is designed for learning from streaming data such as video and is hence based on on-line learning. Additionally, we propose a multi-class co-training approach which allows to also discriminate between the different vehicle classes and against the background. Furthermore, using on-line learning, we do not need to store any data which together with the reduced complexity due to scene-specific training enables our method to be implement on hardware-constrained embedded systems.

2. Audio-Visual Co-Training

In supervised learning one deals with a labeled dataset $\mathcal{D}^L \subseteq \mathcal{X} \times \mathcal{Y} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{|\mathcal{D}^L|}, y_{|\mathcal{D}^L|})\}$, where $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^P$ and $y_i \in \mathcal{Y} = \{+1, -1\}$. In contrast, unsupervised methods aim to find an interesting (natural) structure in $\mathcal{X}$ using only unlabeled input data $\mathcal{D}^U \subseteq \mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathcal{D}^U|}\}$. Co-training [3] is an approach that exploits both labeled $\mathcal{D}^L$ and unlabeled $\mathcal{D}^U$ data.

In co-training, the main idea is that two initial classifiers h_1 and h_2 are trained on labeled data $(\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j) \in \mathcal{D}^L$. Then, these classifiers update each other using the unlabeled data set $\mathcal{D}^U$, if one classifier is confident on a sample whereas the other one is not. The approach has proven to converge [3], if two assumptions hold: (i) the error rate of each classifier is low and (ii) the views are conditionally independent. However, the second condition, which is hard to fulfill in practice, was later relaxed (*e.g.*, [2, 1, 29]). For practical usage, this means that co-training can even be applied, if the learners are only slightly correlated.

Co-training needs two distinct views on the classification problem in order to work. Using audio and video sensors naturally offers "real-world" views, which can be exploited by co-training. Existing co-training approaches used for learning visual classifiers combined different simple cues based on shape, appearance, or motion (*e.g.*, [17, 13, 23, 19]). Thus, starting with Levin *et al.* [17], who indented to train a car detector, co-training was applied for various different applications such as learning a person detector (*e.g.*, [23, 19]), tracking (*e.g.*, [13]), estimating a background model (*e.g.*, [31]).

As it was shown in [29], co-training needs non-coherent views onto the input samples. This can be accomplished by, first, using different sensor sources and, second, by using different classifiers. Additionally, in systems that train from unlabeled data without human supervision, class-label noise can never be avoided. This means that robust co-training classifiers have to be also resistant up to a certain amount of label noise.

In the following, we present a system that uses on-line random naïve Bayes classifiers for the audio data and on-line multi-class boosting based on a logistic loss function for the visual classifier. Both classifiers are inherently multi-class, on-line compatible and noise-robust.

2.1. On-line Random Naïve Bayes

Naïve Bayes classifiers are the simplest kind of Bayesian networks [9] that assume conditional independence of features given a certain class. Thus, the estimation of the unknown joint probability distribution is essentially simplified, i.e., a product of independent likelihoods and a classifier can be formulated as

$$\hat{f}(\mathbf{x}) = \arg\max_i \prod_k p(x_k|c_i), \quad (1)$$

where $p(x_k|c_i)$ is the i^{th} likelihood given class c_i, $p(\mathbf{x}|c_i)$ is the joint probability distribution of feature vector $\mathbf{x}$ given class c_i assuming a uniform class label distribution.

Since single naïve Bayes classifiers are rather weak in accuracy, we build an ensemble of several classifiers similar to [5, 21]. To increase the robustness and stability of bagging [4], we perform random input and random feature selection, which increases the diversity of the classifiers. Using such a random naïve Bayes classifier for visual on-line learning has been recently proposed in [10], where equally binned histograms have been used to estimate the probability distribution for a given feature x_k. We adapt this approach to be used for on-line learning of our audio classifier.

2.2. On-line Multi-class GradientBoost

Boosting [8] additively combines several weak classifiers to a strong one in the form

$$F(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i f_i(\mathbf{x}), \quad (2)$$

where α_i determines the influence of the i^{th} weak learner. During learning, boosting keeps a weight distribution over the training samples. Grabner *et al.* [11] adapted the on-line boosting approach of Oza and Russell [20] to perform on-line feature selection, which is useful in many vision applications that deal with highly overcomplete feature sets.

Since boosting is highly susceptible to class-label noise [18], we adapt the recent approach of Leistner *et al.* [16] which allows the use of robust loss-functions within an on-line gradient boosting framework. To increase the robustness of our classifier, we implement the logit loss $\log(1 + e^{-f_{t,y_t}(\mathbf{x}_t)})$ within the boosting framework. Recently, Saffari *et al.* [26] developed an on-line variant of LPBoost, which is applicable to multi-class classification. For comparison, they also transformed the approach of Zou *et al*. [32] to the on-line domain, which is the same as extending [16] to the multi-class case. Finally, this gives us a robust, multi-class on-line feature selection algorithm which we use for training our visual detector.

2.3. On-line Co-Training System

In off-line co-training, both classifiers are first trained with labeled data. Subsequently, all unlabeled samples are evaluated by both classifiers and ranked by their confidences. Confident samples of classifier A are then integrated into the labeled training set of classifier B and vice versa. This procedure is repeated several times. In the online domain, we have to evaluate each sample $\mathbf{x}_t$ individually. If one classifier makes a confident decision on a sample, it predicts a pseudo label $\hat{y}_1^t$ and updates the second one with the according view on the data.

Algorithm 2.1 On-line Audio-Visual Co-Training

```
Require: classifiers $F_1^0$ and $F_2^0$
Require: labeled data $\{\mathbf{x}, y\}^N$
Require: unlabeled data $\{\mathbf{x}\}^M$
Train classifiers on labeled data $\{\mathbf{x}, y\}^N$
for each unlabeled input sample $\mathbf{x}_t \in \{\mathbf{x}\}^M$ do
    // Evaluate classifiers and estimate pseudo labels $\hat{y}_{1,2}^t$
    $\hat{y}_1^t \leftarrow$ eval($F_1^{t-1}, \mathbf{x}_t$)
    $\hat{y}_2^t \leftarrow$ eval($F_2^{t-1}, \mathbf{x}_t$)

    // On-line Learning
    $F_1^t \leftarrow$ update($F_1^{t-1}, \mathbf{x}_t, \hat{y}_2^t$)
    $F_2^t \leftarrow$ update($F_2^{t-1}, \mathbf{x}_t, \hat{y}_1^t$)

    // Classifiers are available anytime
    Output: classifiers $F_1^t$ and $F_2^t$
end for
```

2.4. Classifier Synchronization

Our classifiers work on different scopes regarding capture time and object localization (*i.e.*, still images for visual classification and audio streams of several seconds length for audio classification). Thus, we have to synchronize both, the evaluation and the co-training of our classifiers. This is accomplished by using a visual trigger that robustly delivers points in time, where vehicles are present within a limited region in the visual view. The visual trigger uses a robust block-based background model [22], which is able to handle camera shake caused by wind and vibrations and permits changing illumination conditions due to on-line adaption of the model. The background model captures the mean intensity of the background within small rectangular blocks and signals foreground objects at positions where the deviation of the mean from one frame to another is larger than a threshold. We define a small region of interest spanning over all lanes for our visual triggering and report a passing vehicle if some foreground object is detected within this region (see Figure 1 for two illustrative examples).

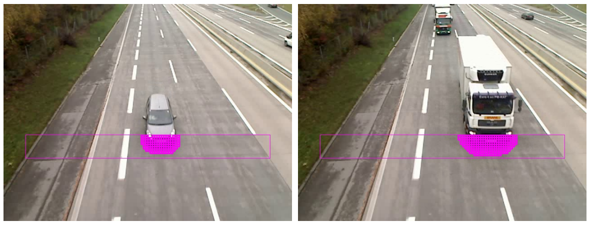

(a) (b)

Figure 1. Visual trigger

The visual classifier then densely evaluates this region and delivers detections of present vehicles. Since the acoustic features are calculated on an extended time range, we have to process the audio stream 3 seconds forward and backward in time. Both classifiers then use their classification result $\hat{y}^t$ to train the other one. Thus, the visual classifier uses the location that results in the highest confidence measure according to the class $\hat{y}^t$ delivered by the audio classifier.

2.5. Evaluation

During the classification phase, we unify the visual and audio cue by linearly combining the confidences of both classifier types. To classify a scene, we first generate candidate regions for both classes, cars and trucks by applying our visual classifier. We then combine the visual classifiers' confidences with the confidences provided by the audio classifier. (All confidences are normalized to the range of [-1, +1] before fusion.) 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, both set to 0.5 (better values for α and β could be found by using cross-correlation on labeled samples or other weighting techniques.). Finally, by using a non-maxima suppression, the highest vote is estimated by providing the according class for the candidate regions.

2.6. Features

In this section, we describe the features we extract on our platform. Both, visual and acoustic features can be computed with minimal computational requirements and in real-time.

Acoustic Features In general, it is a non-trivial task to end up with a set of robust acoustic features due to the non-stationarity of the audio signals. Noise inferences such as bypassing vehicles on neighboring lanes on highways are additionally distorting the captured signal. We use a set of block-based acoustic features recently proposed by Starzacher *et al.* [27] which have shown to be suitable for efficient implementation on our embedded platform and highly class-discriminative. The selected acoustic features are short-time energy, spectral roll-off point, spectral bandwidth, band energy ratio values and mean cepstral coefficient. Classifiers exclusively based on these features have shown to achieve classification accuracy above 90%.

Visual Features Due to the constrained settings of our platform, *i.e.*, limited memory, computational power and visual resolution of the camera, Haar-like features [28] have shown to represent a reasonable trade off. We use an extended set of 6 different configurations representing edges, lines and radial structures. For training our classifier, we use a randomly initialized feature pool. Due to the large number of classifiers within our classifier bag, a suitable model can be trained. Additionally, to describe coarse structures of our object we use Haar-LBP Features [24]. Both feature types can be computed very fast and efficient using integral structures.

3. Experimental Evaluation

To demonstrate the performance of our approach, we conduct several experiments evaluating the co-training. We have recorded training data on different locations and performed training labeled and unlabeled training.

3.1. Data and Study Locations

The audio-visual co-training framework is applied to real-world datasets recorded on several different locations under varying weather conditions (specified as $Data_1$ and $Data_2$). The vehicles of interest are cars and trucks. Basically, a consumer microphone and camera were placed on a bridgeover to record passing vehicles. Data_1 consists of about 500 cars and 300 trucks, whereas Data_2 obtains approximately 160 car and 90 truck samples. The audio recording was performed with 8 kHz sampling rate, 16bit resolution, mono format and an average maximum recording duration of 5 seconds. The frame rate of the camera was set to approximately 20 Hz with a resolution of 640 × 480.

3.2. Embedded Test Platform

A MicroSpace EBX (*MSEBX945*) embedded computer board from *DigitalLogic* serves as our evaluation platform. It offers a compact EBX single-board construction of 146mm × 203mm and provides several interfaces such as RS-232, FireWire, USB and LAN 100 Mbit/s. The *SMX945-L7400* CPU module is based on an Intel Core 2 Duo processor with 2 × 1500 MHz and a 667 MHz FSB. It runs *Linux from Scratch* and the total power consumption of the platform is about 12 to 15 W.

Cars	Recall	Precision	F-Measure
50 Lab. Samples	0.86	0.85	0.85
150 Lab. Samples	0.98	0.97	0.98
200 Lab. Samples	0.96	0.96	0.96
Trucks	Recall	Precision	F-Measure
50 Lab. Samples	0.65	0.48	0.45
150 Lab. Samples	0.89	0.88	0.88
200 Lab. Samples	0.82	0.94	0.88

Table 1. Initial performance using different amounts of labeled Data.

Cars	Recall	Precision	F-Measure
Initial (100 Lab.)	0.96	0.94	0.95
Initial + 100 Unl.	0.94	0.92	0.93
Initial + 200 Unl.	0.96	0.93	0.94
Trucks	Recall	Precision	F-Measure
Initial (100 Lab.)	0.75	0.58	0.65
Initial + 100 Unl.	0.86	0.85	0.86
Initial + 200 Unl.	0.91	0.80	0.85

Table 2. Co-training performance using different amounts of unlabeled Data.

3.3. Baseline performance

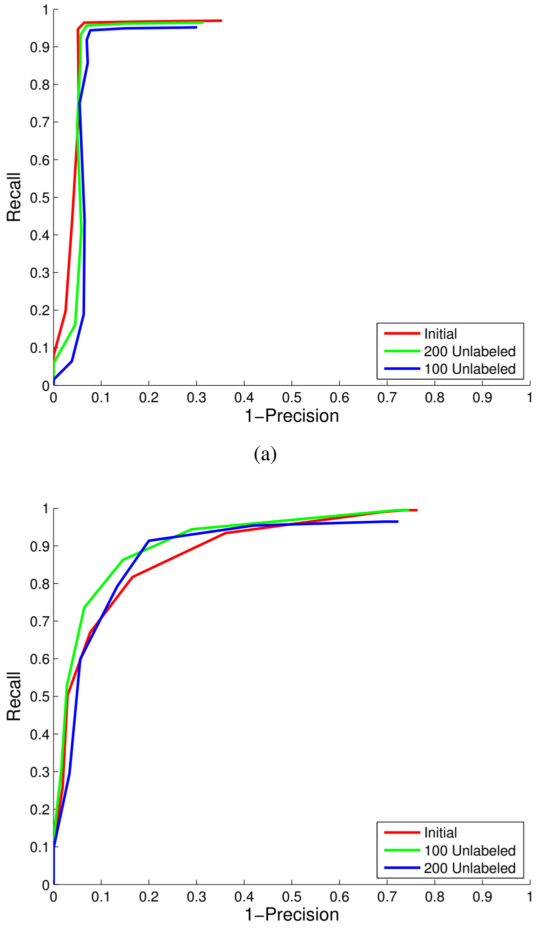
To have a fair comparison, we have to first perform an evaluation only using labeled data for training. As expected, the performance for car localization already performs well after only 100 samples, where truck localization only yields poor performance. Table 1 depicts detailed results. If we are using only 50 labeled samples, the visual classifier is not able to cope with the various truck appearances and only delivers very poor performance.

3.4. Co-Training Convergence

As our aim is to gain the same performance by using only a few labeled training data, we use only 100 labeled samples (initial classifier) and perform co-training [17] subsequently. Figure 2 and Table 2 show that we can achieve a huge performance improvement for trucks by performing audio-visual co-training without changing the already sufficient performance for car classification. Of course, the performance for purely supervised training is higher, but therefore we would require a larger amount of labeled examples.

3.5. Labeled vs. Unlabeled Samples

An interesting aspect of our approach is the influence of the amount of labeled samples to the performance of the fi-



(a)

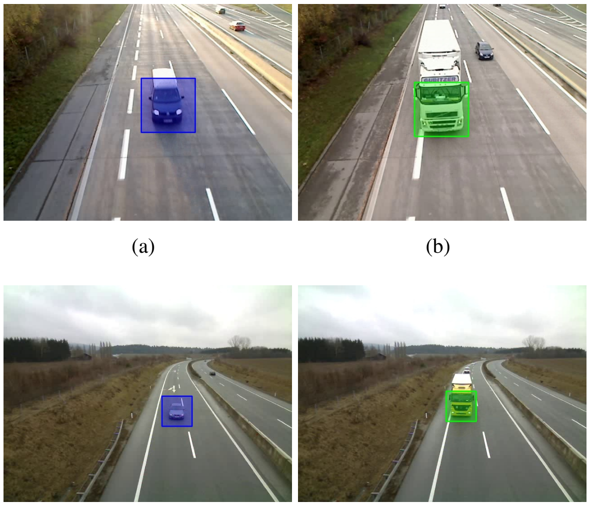
(b)

Figure 2. Co-training performance for (a) cars and (b) trucks.

nal classifier. Therefore, we use an amount of 200 samples to train the classifier, where we vary the number of labeled samples between 50 and 150 and use the remaining samples as unlabeled. It is clearly visible (see Table 3), that we can improve the classifiers performance with unlabeled data once a certain number of labeled samples has been used and both classifiers reach a reasonable performance. If we are only using a very small amount of labeled samples to train the initial classifier (*e.g.*, 50), we violate one of the co-training assumptions since then the error rate of the visual classifier is too high. Performing co-training with such a weak classifier runs into the risk of degrading the performance of both classifiers due to the large amount of noise (*i.e.*, false predictions $\hat{y}^t$) introduced in the training process. In our case, the performance does not degrade only due to the stability of the random naïve Bayes classifier used for acoustic classification.

Cars	Recall	Precision	F-Measure
50 Lab. + 150 Unl.	0.99	0.97	0.98
100 Lab. + 100 Unl.	0.94	0.92	0.93
150 Lab. + 50 Unl.	0.98	0.97	0.98
Trucks	**Recall**	**Precision**	**F-Measure**
50 Lab. + 150 Unl.	0.67	0.82	0.74
100 Lab. + 100 Unl.	0.86	0.85	0.86
150 Lab. + 50 Unl.	0.91	0.80	0.85

Table 3. Classifier Performance using different amounts of labeled and unlabeled training data.



(a)

(b)

(c)

(d)

Figure 3. Snapshots with groundtruth from dataset (a, b) Data$_1$ and (c, d) Data$_2$. It is clearly visible that both locations are captured with different viewing angle and scale.

3.6. Scene Adaption

Usually when performing visual vehicle detection without a priori knowing where the site of operation will be, a large and complex classifier is trained to discriminate the target object from every possible background. Since we are now able to improve our classifier using unlabeled data, this experiment shows that we can also adapt an already trained vehicle classifier to a new location. Therefore, we train our classifier initially with labeled positive data from a scene (300 labeled sample per class from $Data_2$). Subsequently, we use cropped background samples from our new scene ($Data_1$) for bootstrapping and adapt the classifier during runtime by using samples generated by our co-training approach.

Since both scenes have different camera orientations (see Figure 3), we resize our visual classifier to satisfy the new geometry constraints. For the audio classification, the changed environment only affects the short-time energy feature, which was simply removed from the used feature set

Cars	Recall	Precision	F-Measure
Initial (100 Lab.)	0.99	0.94	0.96
Initial + 100 Unl.	0.99	0.97	0.98
Initial + 200 Unl.	0.99	0.99	0.99
Trucks	Recall	Precision	F-Measure
Initial (100 Lab.)	0.48	0.46	0.47
Initial + 100 Unl.	0.63	0.82	0.71
Initial + 200 Unl.	0.70	0.71	0.70

Table 4. Classifier performance for scene adaption after 100 and 200 unlabeled samples from the target scene

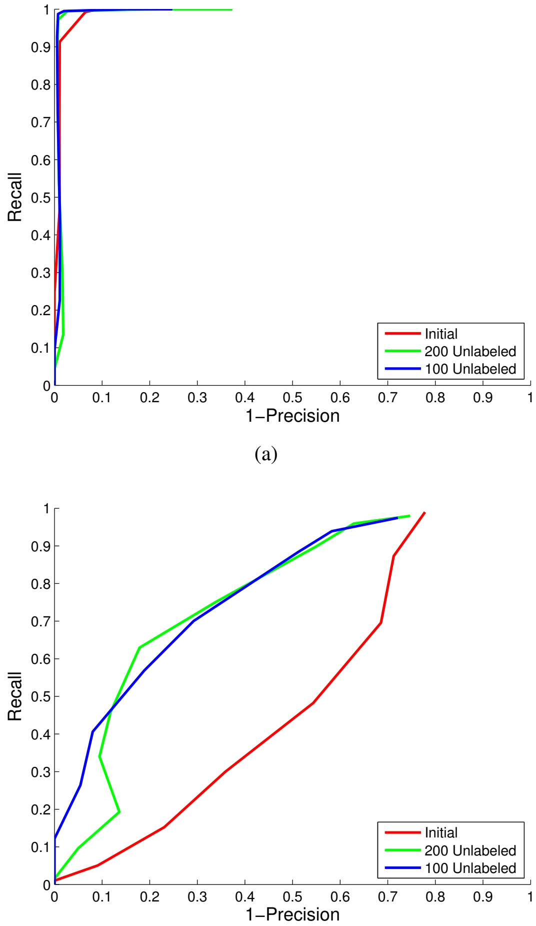
Cars	Recall	Precision	F-Measure
Audio-Visual	0.96	0.93	0.94
Visual-Only	0.77	0.64	0.70
Trucks	Recall	Precision	F-Measure
Audio-Visual	0.86	0.85	0.86
Visual-Only	0.75	0.85	0.80

Table 5. Co-training performance using audio-visual and visual-only classifier combinations.

for this experiment. This decreases the performance of the acoustic classification only by a few percent. We report the classifier performance after training our classifier with 100 and 200 unlabeled samples from the new scene. Table 4 depicts the improvement in classification accuracy, by only using such a small amount of unlabeled data. Even if the performance for car classification is already quite good, the co-training process improves the results further. Also for truck classification a clear improvement in accuracy can be gained. Since these samples can be generated during runtime for free, this is an easy way to adapt an already trained classifier to a new task not having even a single labeled sample from the new scene.

3.7. Homogeneous vs. heterogeneous Co-training

Finally, we want to compare homogeneous and heterogeneous co-training to further emphasize the applicability of our approach. Therefore, we compare our co-training approach using acoustic and visual classifiers against co-training using two visual classifiers selecting features out of two independent randomized feature pools. We train all classifiers using 100 labeled training samples and subsequently perform audio-visual and visual-visual co-training respectively. Table 5 depicts the evaluation results comparing visual-only and audio-visual co-training after 100 labeled plus 200 unlabeled samples. It is clearly visible, that the visual-only co-training cannot match the performance of the mixed approach due to the smaller diversity among the two visual classifiers.



(a)

(b)

Figure 4. Classifier performance for scene adaption for (a) cars and (b) trucks.

4. Conclusions

In this paper, we introduced a novel on-line co-training method based on audio and video information. Our system uses simple, robust and rapid classifiers in order to be implemented on hardware-constrained embedded platforms. In order to ensure easy deployment and maintenance, we do not use complicated calibrations or microphone arrays; instead, all of our system components are off-the-shelf consumer products. We propose to use an on-line multi-class gradient boosting for visual classification which inherently allows to discriminate between cars and trucks as well as the local background. For increased diversity between audio and visual classification, we use an on-line random naïve Bayes classifier for acoustic classification. In the experiments, we demonstrated that our system robustly adapts to traffic scenes without using any additional human labeling effort. In our *Scene Adaption Experiment*, we clearly show that the better the performance of the initial classifier is, the lower is the number of false predicted labels for the co-training process and the better the co-training works. Our work shows that even for a very small amount of labeled data, we can gain reasonable classification accuracy by combining heterogeneous classifiers and improve the classifier during runtime by on-line co-training. In future work, we will try to exploit a significant larger amount of unlabeled data. Additionally, we will focus on more suitable representations for trucks in order to improve the overall detection results.

References

[1] S. Abney. Bootstrapping. In *Proc. Annual Meeting of the Association for Computational Linguistics*, 2002. 2

[2] M.-F. Balcan, A. Blum, and K. Yang. Co-training and expansion: Towards bridging theory and practice. In *Advances in Neural Information Processing Systems*, 2004. 2

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. Conf. on Computational Learning Theory*, 1998. 2

[4] L. Breiman. Bagging predictors. *Machine Learning*, 1996. 2

[5] L. Breiman. Random forests. *Machine Learning*, 2001. 2

[6] C. Christoudias, R. Urtasun, A. Kapoorz, and T. Darrell. Co-training with noisy perceptual observations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. 1, 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. 1

[8] Y. Freund and R. Schapire. A short introduction to boosting. *Journal of the Japanese Society for Artificial Intelligence*, 1999. 2

[9] D. Geiger, M. Goldszmidt, G. Provan, P. Langley, and P. Smyth. Bayesian Network Classifiers. In *Machine Learning*, 1997. 2

[10] M. Godec, C. Leistner, A. Saffari, and H. Bischof. Online random naive bayes for tracking. In *Proc. Intern. Conf. on Pattern Recognition*, 2010. 2

[11] H. Grabner and H. Bischof. On-line boosting and vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006. 2

[12] H. Hu and J. Q. Gan. Sensors and Data Fusion Algorithms in Mobile Robotics. *Technical Report: CSM-422, Department of Computer Science, University of Essex, United Kingdom*, 2005. 1

[13] O. Javed, S. Ali, and M. Shah. Online detection and classification of moving objects using progressively improving detectors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2005. 2

[14] A. Klausner, S. Erb, and B. Rinner. DSP Based Acoustic Vehicle Classification for Multi-Sensor Real-Time Traffic Surveillance. In *Proc. European Signal Processing Conference*, 2007. 1

[15] M. Kushwaha, S. Oh, I. Amundson, X. Koutsoukos, and A. Ledeczi. Multi-Modal Target Tracking Using Heterogeneous Sensor Networks. In *Proc. Intern. Conf. on Computer Communications and Networks*, 2008. 1

[16] C. Leistner, A. Saffari A. A., P. M. Roth, and H. Bischof. On robustness of on-line boosting - a competitive study. In *Proc. IEEE On-line Learning for Computer Vision Workshop*, 2009. 2, 3

[17] A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2003. 2, 4

[18] P. M. Long and R. A. Servedio. Random classification noise defeats all convex potential boosters. In *Proc. Intern. Conf. on Machine Learning*, 2008. 2

[19] V. Nair and J. J. Clark. An unsupervised, online learning framework for moving object detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2004. 2

[20] N. C. Oza. *Online Ensemble Learning*. PhD thesis, University of California, Berkeley, 2001. 2

[21] A. Prinzie and D. V. den Poel. Random multiclass classification: Generalizing random forests to random mnl and nb. In *Database and Expert Systems Applications*, 2007. 2

[22] P. M. Roth. *On-line Conservative Learning*. PhD thesis, Graz University of Technology, Faculty of Computer Science, 2008. 3

[23] P. M. Roth, H. Grabner, D. Skočaj, H. Bischof, and A. Leonardis. On-line conservative learning for person detection. In *Proc. IEEE Workshop on VS-PETS*, 2005. 2

[24] A. Roy and S. Marcel. Haar local binary pattern feature for fast illumination invariant face detection. In *Proc. British Machine Vision Conf.*, 2009. 4

[25] H. Ruser and F. P. León. Informationsfusion - Eine Übersicht. *tm - Technische Messen, Oldenburg*, 2007. 1

[26] A. Saffari, M. Godec, T. Pock, C. Leistner, and H. Bischof. On-line multi-class lpboost. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010. 2

[27] A. Starzacher and B. Rinner. Single Sensor Acoustic Feature Extraction for Embedded Realtime Vehicle Classification. In *Proc. Intern. Workshop on Sensor Networks and Ambient Intelligence*, 2009. 4

[28] P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001. 1, 4

[29] W. Wang and Z.-H. Zhou. Analyzing co-training style algorithms. In *Proc. European Conf. on Machine Learning*, 2007. 2

[30] W. Zajdel, J. D. Krijnders, T. Andringa, and D. M. Gavrila. CASSANDRA: audio-video sensor fusion for aggression detection. In *Proc. IEEE Intern. Conf. on Advanced Video and Signal based Surveillance*, 2007. 1

[31] Q. Zhu, S. Avidan, and K.-T. Cheng. Learning a sparse, corner-based representation for background modelling. In *Proc. IEEE Intern. Conf. on Computer Vision*, 2005. 2

[32] H. Zou, J. Zhu, and T. Hastie. New multi-category boosting algorithms based on multicategory fisher-consistent losses. *The Annals of Applied Statistics*, 2008. 3