Scene Segmentation in Adverse Vision Conditions

Evgeny Levinkov*

Max Planck Institute for Informatics, Saarbrücken, Germany
levinkov@mpi-inf.mpg.de

Abstract. Semantic road labeling is a key component of systems that aim at assisted or even autonomous driving. Considering that such systems continuously operate in the real-world, unforeseen conditions not represented in any conceivable training procedure are likely to occur on a regular basis. In order to equip systems with the ability to cope with such situations, we would like to enable adaptation to such new situations and conditions at runtime. We study the effect of changing test conditions on scene labeling methods based on a new diverse street scene dataset. We propose a novel approach that can operate in such conditions and is based on a sequential Bayesian model update in order to robustly integrate the arriving images into the adapting procedure.

1 Introduction

Driving assistance systems have been rapidly evolving lately due to a constantly increasing interest in real-world application as well as studies conducted in the field of computer vision. An important task of such systems is road scene labeling in order to derive the semantic structure of the observed scenes. One of the big challenges is making such systems robust so that they can reliably operate in a wide range of conditions. However, capturing and training every possible condition a car can encounter throughout years of driving seems to be an impossible task.

Recently, there has been an increased interest in approaches of domain adaptation [8, 7] in computer vision that are able to adapt existing classifiers to new domains and conditions. These require supervision from the target domain, that can not be provided by the envisioned systems that continuously operate in the real-world. Existing adaptive methods [1] allow the use of machine generated labels in order to refine the classifier and help it to adapt to changing conditions. However, they perform only global adaptation, for which they require access to the whole test set. Again, this is against the idea of a continuously operating system.

In contrast, we aim at an adaptive algorithm that is able to perform adaptation on the fly. Therefore, this paper proposes a sequential Bayesian update strategy that pursues multiple model hypothesis for semantic scene labeling. In

* Recommended for submission to YRF2014 by Dr. Mario Fritz

order to circumvent typical problems of online learning by a "self-training" procedure, we perform model updates under the assumption of a stationary label distribution.

2 Naïve Model Update

Typical self-training approaches are based on a two step procedure. First, a lately arrived batch of images is labeled using the current model. Second, after an optional threshold on a confidence rating, these samples are used to update/re-train the model. In more detail, we get an output probability distribution $P(x_{(i,j)})$ from our classifier for each pixel (i,j) and the predicted class-label for it $c^* = \text{argmax}_{c \in \mathcal{Y}} P(x_{(i,j)} = c)$. Then, samples are taken for which $P(x_{(i,j)} = c^*) > \lambda$ holds, where λ is a acceptance threshold parameter. High probability $P(x_{(i,j)} = c^*)$ should indicate high confidence of the classifier in the predicted label. This is a completely heuristic approach, as the classification of the test data is only an approximation to the un-accessible groundtruth.

Taking new samples with the predicted labels which have high confidence is not necessarily a reliable way of updating the model due to inaccuracies in the intermediate models. While we want to be robust w.r.t. changes in the feature distribution, stationarity of the label distribution is a milder assumptions in many scenarios. We adopt ideas from J. Alvarez *et al.* [1] who employ a pixel-wise, normalized class-histogram on the off-line data as a prior distribution to weight the output probability distribution of the classifier at testing time.

In detail, we compute histogram for each pixel and after per-pixel L_1- normalization we get a prior $P_{pr}^{(i,j)}$ for each pixel $(i,j), i = 1, \ldots, W_{pr}, j = 1, \ldots, H_{pr}$. In our experiments images in the testing dataset all have various dimensions, so we perform nearest-neighbor sampling from the prior distribution $P_{pr}^{(i,j)}$. Then at testing time output probability distribution $P(x_{(i,j)})$ for all pixels $(i,j), i = 1, \ldots, W, j = 1, \ldots, H$ from our classifier for an image with dimensions $W \times H$ is element-wised multiplied with the corresponding prior

$$\tilde{P}(x_{(i,j)}) \propto P(x_{(i,j)}) P_{pr}^{(\lfloor i \frac{H_{pr}}{H} \rfloor, \lfloor j \frac{W_{pr}}{W} \rfloor)}. \tag{1}$$

This is used for accepting or rejecting new training examples on a per-pixel basis $\tilde{P}(x_{(i,j)} = c^*) > \lambda$.

3 Sequential Bayesian Model Update under Structured Scene Prior

We propose a new model to leverage unlabeled data for a sequential model update for scene labeling. Our approach is based on a Bayesian model update. We maintain a population of models (particles) that approximate the distribution over the model-space $p(h_t|L_t)$, instead of relying on a single model, as in the previous formulations. The required integration over the model-space is solved

by a Monte-Carlo method – just like in Condensation and Particle Filters that are well known from tracking applications [5, 4]. Consequently, scene labeling at test time will be performed by marginalization over the model distribution

$$p(X|L_t) = \int p(X|h_t)p(h_t|L_t)\,\mathrm{d}h_t, \tag{2}$$

where X is the labeling of a test image for which we want to do prediction.

While the above-mentioned tracking formulations have a measurement step that evaluates image evidence, we measure the compatibility with the scene prior S. This is again based on the assumption of a stationary label distribution $P_{pr}^{(i,j)}$ as for the previous method.

Bayesian Model Update We are interested in modeling an evolving target distribution over models in order to account for the uncertainty in the unobserved scene labels. Therefore, we model the unobserved scene labels l_t of the unlabeled data u_t at time step t as a latent variable (Figure 1). Rather than sticking to a single model hypothesis, we seek to model a distribution over model hypothesis h_t. Therefore we update a distribution over model hypothesis given labels $p(h_t|L_t)$. Here $L_t = \{l_0, l_1, \ldots, l_{t-1}, l_t\}$.

We describe the incorporation of the unlabeled examples in a Bayesian framework by integrating over all model hypothesis

$$p(h_t|L_{t-1}) = \int p(h_t|h_{t-1}, u_t)p(h_{t-1}|L_{t-1})\,\mathrm{d}h_{t-1}. \tag{3}$$

In the measurement step, we apply the Bayes' rule in order to get the updated distribution over model hypothesis

$$p(h_t|L_t) = \frac{p(l_t|h_{t-1}, S)p(h_t|L_{t-1})}{p(l_t|L_{t-1})}, \tag{4}$$

with

$$p(l_t|h_{t-1}, S) = p(l_t|h_{t-1})p(l_t|S), \tag{5}$$

where $p(l_t|h_{t-1})$ is the probability of a certain scene labeling prediction given a model hypothesis h_{t-1} and $p(l_t|S)$ is a scene labeling prior $P_{pr}^{(i,j)}$.

Sampling We perform inference with a Monte-Carlo approach [5]. At each time step the model distribution $p(h_t|L_t)$ is represented by a set of particles $s_t^{(N)}$ with weights $\pi_t^{(N)}$. Next, the particles are propagated to the next time step via $p(h_t|h_{t-1}, u_t)$ that takes into account the existing models and the unlabeled data. In traditional tracking application this transition is modeled with a deterministic part and a stochastic component. In our setting, we propose to do model propagation by randomly choosing a subset of images which are provided to a particular particle to retrain as well as picking a randomized acceptance threshold λ per particle. The benefits are twofold. First, a diverse set of models is generated for the next iteration. Second, parameters like the acceptance thresholds are dealt with within the model and no hard choices have to be made.

Fig. 2. First column shows examples of road scene dataset from [9]. Other columns show examples of the new diverse road scene dataset exhibiting very different appearances and a wider range of conditions.

In summary, our particle filter over model space works as follows. For each particle i out of N:

1. Pick a particle s_t^i from $s_t^{(N)}$, which represents $p(h_t|L_t)$, according to the weights $\pi_t^{(N)}$
2. Sub-sample set of unlabeled images u_t to $\hat{u}_t$
3. Predict labels $\hat{l}_t = \text{argmax}_l\, p(l|h_t)$ for subset $\hat{u}_t$
4. Accept or reject samples based on some threshold λ
5. Retrain model using $(\hat{u}_t, \hat{l}_t)$ and L_{t-1}

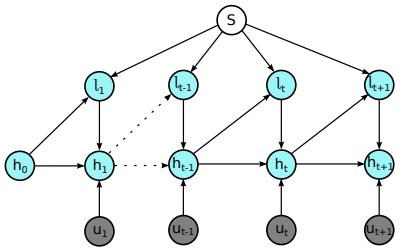

Fig. 1. Bayesian network for the proposed Sequential Bayesian model update.

Traditional tracking approaches would now follow up with a measurement in order to update the weights $\pi_t^{(N)}$. Similarly, we update the weight $\pi_t^{(N)}$ of each sample (model hypothesis) according to (4). In this equation $p(h_t|L_{t-1})$ is the distribution represented by our particles after the propagation step from above and $p(l_t|h_{t-1}, S)$ is the product of the likelihood of the labeling times the likelihood of the labeling given the scene labeling prior. We don't compute the denominator - but rather directly normalize the weights of the particles $\pi_t^{(N)}$ to sum to 1.

4 New Diverse Road Scenes Dataset

In order to study the problem of adaptation we need a dataset, which exhibits considerable amount of appearance variation between the training and test set. Typical road scene datasets like [9, 2] (Figure 2, first column) already exhibit some visually difficult situations like changes in object appearances due to motion blur effect, deep shadows which appear and disappear suddenly, changes in lightning conditions like over- or under-saturated regions, but the overall feature

Test set	Fully connected CRF error, %			
	Road	Background	Sky	Average
Old	0.7	2.2	2.7	1.9
New	52.7	6.5	35	31.4

Table 1. Comparison of Krähenbühl *et al.* [6] semantic image labeling algorithm on the old and the new test test.

statistics stays similar between training and test. Therefore, we have collected a new dataset which exhibits much richer appearance variation, using freely available images from the Internet. Figure 2 shows examples from 220 images in our dataset. All images were hand-labeled into 3 classes: road, sky, and background. The dataset expose a much stronger appearance variation than previous datasets. Typical challenges include roads covered with autumn leaves or snow as well as different types of roads such as dirt and gravel roads and even images taken at night, although we leave out such issues like bad lighting, low contrast, or rain.

5 Experimental Results

In our implementations we employed a Random Forest [3] classifier (consisting of 10 trees each having maximal depth 15 with 20% bagging) and features from [9]. We used the training set from [9] for training (Figure 2, first column) and performed testing or adaptation on the new diverse road test set. Groundtruth annotation of the test set is not used in any way, other than for computing error rates.

Non-adaptive methods In order to show that non-adaptive methods have a limited capability of generalizing to a different and strongly varying feature distribution as presented in our new dataset, we took one of the state-of-the-art methods for semantic image labeling of Krähenbühl *et al.* [6], and trained it on the training set and tested on both the old and the new test set (Table 1). The old test set has a similar appearance as the training set (Figure 2, first column), so the resulting numbers are very good. But when we test on the new test set, the method shows strong accuracy degradations caused by the changed feature distribution. Particularly, the road recognition rate gets more than 50 times worse, because background and sky have more or less similar appearance as in the training set, while appearance of the road usually does not resemble the one in the training set.

Adaptive methods Global adaptive methods consider the whole test set at once and try to adapt to it. The main restriction of such methods is that they require access to the whole test set. In the real world setting, when new images constantly arrive, global algorithms would have to deal with a constantly increasing test set. Alvarez *et al.* [1] proposed such an globally adaptive scheme for road scene segmentation. Table 2 (first row) presents resulting numbers for their original method, which the authors kindly agreed to run on our test set.

Table 2 shows resulting numbers for adaptive methods after they have processed the whole test set. Our algorithms were run 3 times and the results were

Update type	Method	Error, %			
		Road	Background	Sky	Average
global	Alvarez *et al.* [1]	76.2	**12.7**	25.5	38.2
sequential	Naïve	26 ±1.4	15.4±0.4	9.3±1.4	17±0.7
	Naïve + Scene Prior	21±2.7	18.5±0.6	6.5±0.9	15.5±1.4
	Bayesian Model	**19±0.6**	18.3±0.6	**4.5±0.4**	**13.9±0.3**

Table 2. Comparison of different adaptive approaches after processing the whole test set (mean plus std). Bold font highlights the best numbers.

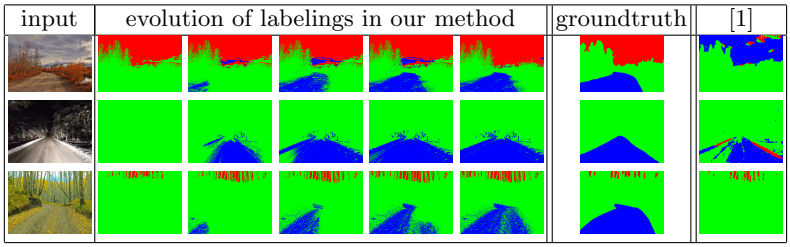

Fig. 3. Example results showing the input image and evolution of the labelings through the proposed Sequential Bayesian Update method. The last two columns show the corresponding ground truth annotation and the output of the global adaptive method of Alvarez *et al.* [1]. Green color denotes background, red - sky, and blue - road.

averaged over. The numbers show that sequential adaptive methods have a larger capability in adapting to changing feature distribution in the setting, when new images arrive constantly during (possibly infinite) test time. Our proposed Sequential Bayesian Model Update shows the best average performance and the lowest variance.

Figure 3 shows some examples of how labelings for certain images evolve as our Bayesian Model Update method processes one batch of consequent images from the test set after another. It is remarkable how our approach can recover from initially poor segmentation results and adapts to the new conditions. We also show the results of the method of Alvarez *et al.* [1], over which we show quantitative as well as qualitative improvements.

6 Conclusion

Today's semantic scene labeling methods show good performance if the training distribution is representative for the test scenario. But when this feature distribution does change, as we showed, such techniques deteriorate in performance quickly. We collected a challenging dataset of images which has very different appearance statistic compared to the established scene segmentation datasets.

We proposed a Bayesian Model Update that sequentially updates the segmentation model as new data arrives, allowing to benefit from new information at test time and providing a possible application in scenarios when the new data is not available all at once, but rather arrives constantly in small batches.

References

1. Álvarez, J.M., Gevers, T., LeCun, Y., López:, A.M.: Road scene segmentation from a single image. In: ECCV (2012)
2. Álvarez, J.M., López, A.M.: Road detection based on illuminant invariance. In: ITS (2010)
3. Breiman, L.: Random forests (2001)
4. Dellaert, F., Burgard, W., Fox, D., Thrun, S.: Using the condensation algorithm for robust, vision-based mobile robot localization. In: CVPR (1999)
5. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. In: IJCV (1998)
6. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS (2011)
7. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: CVPR (2011)
8. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV (2010)
9. Wojek, C., Schiele, B.: A dynamic conditional random field model for joint labeling of object and scene classes. In: ECCV (2008)