Scene recognition under special traffic conditions based on deep multi-task learning

Xiaochang Hu, Xin Xu, Yongqian Xiao, Hongjun Chen, Hongia Zhang

1 College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, People’s Republic of China
E-mail: h_zhang00@163.com

Abstract: Traffic scene recognition under special conditions is one of the most promising yet challenging tasks for autonomous driving systems. This study presents a deep multi-task classification framework for scene recognition involving special traffic conditions. The framework incorporates four learning tasks where the recognition of special traffic scenes is the chief task and the time of occurrence (daytime or night-time), the weather type and the road attribute are the three auxiliary tasks for improving the recognition performance. The four tasks share the feature map generated by a convolutional neural network followed by task-specific sub-networks which are merged in the end via a joint loss function. Moreover, a small dataset of typical special traffic conditions was built for training and testing the recognition model. Experimental results demonstrate that the proposed framework significantly improves the accuracy of scene recognition under special traffic conditions.

1 Introduction

It is well known that a practical autonomous driving system requires reliable and effective traffic scene understanding [1–4]. Among them, the recognition of special traffic scenes is vital to the perception model. Not only is it related to the specific driving mission of the autonomous vehicle (e.g. passing through regulation of traffic), but also the safety of the vehicle itself as well as other participants in the traffic is dependent on this.

Nevertheless, special traffic scene recognition is often neglected, and limited effort has been made to explore special traffic scene recognition task so far. A model-based approach such as traffic sign detection was used in papers [5–7] to tackle this problem, but not all special traffic scenes have traffic signs, hence this method has low feasibility. In general, traffic warning signs should be placed when a traffic accident happens as shown in Fig. 1a, but this is not always the case due to negligence or special occasions such as road construction or temporary inspection of the road (Fig. 1b). In the latter case, rule-based traffic sign detection is not reliable and may cause risks since the model-based feature extraction process discards a large amount of scene information and is difficult to cope with uncertain scenes. Some others focus on natural traffic scene recognition [1–4, 8, 9] such as weather type and the time of occurrence. Most of them aim to develop a more robust model with complex scene changes for low-level perception tasks, such as dynamic vehicle detection under different weather/illumination conditions.

Besides algorithm developments, there are also various benchmark datasets [10–15] for automatic driving, but only a few of them include special traffic scene recognition. On the other hand, as the occurrence of special traffic scenes is unpredictable, it is difficult to collect enough data for special traffic scenes.

In this paper, we propose a deep multi-task classification framework where the recognition of special traffic scenes is the chief task, and three auxiliary tasks are designed which include the time of occurrence (daytime or night-time), the weather type and the road attribute, respectively. This framework features the following two main advantages. On the one hand, the multi-task framework is highly favourable since it improves the performance of the chief task by exploiting the intrinsic relationship among the four tasks [16, 17]. On the other hand, based on the end-to-end method of deep learning, the raw scene information is embedded into the feature space and the loss of partial scene information during feature extraction is avoided as a result [5–8]. Aside from the proposed framework, a small dataset of special traffic conditions is built and used to test the proposed framework.

The remainder of this paper contains six sections. Section 2 reviews the related work. Section 3 is the description of the method concerning the architecture of the network and the joint loss function. Section 4 elaborates the data acquisition methods and the structure of dataset, followed by Section 5 where experiments were conducted to evaluate the method. Finally, conclusions are drawn in Section 6.

2 Related work

2.1 Scene recognition

In the past few years, scene understanding has made a great progress. However, most of the effort focuses on weather impact and illumination changes [2–4, 8, 9]. Di et al. [2] proposed a dense correspondence-based transfer learning approach to understand the traffic scene from images taken at the same location but under different weather or illumination conditions. Lu et al. [3] proposed a co-training approach for labelling as either sunny or cloudy by segmentation of the sky followed by determining the weather type. Lu et al. [4] adopted an approach based on an embedded semantic space to solve the problem by adopting a multi-task learning method to solve the semantic gap. Inspired by human scene understanding based on object knowledge, Liao et al. [8] addressed the problem of scene classification by encouraging deep neural networks to incorporate object-level information. Lee and Kim [9] proposed an approach to estimate fog level via intensity curves with geometrical information and a neural network was used to classify features into four fog levels. Most of the existing literature puts effort on weather and illumination and aims to improve the robustness of low-level visual tasks, i.e. target detection and lane

Fig. 1 Two traffic accident scenes with respect to traffic warning sign
(a) With traffic warning sign, (b) Without traffic warning sign
tracking. Very little work focuses on the role of high-level semantic information of the scene in high-level driving tasks and its relationship with behavioural decision-making.

2.2 Multi-task learning

Multi-task learning aims to improve the efficiency of learning and the accuracy of prediction by learning multiple objectives from a shared representation [17]. In computer vision, multi-task learning methods have been studied. The UberNet [18] learned several different regression and classification tasks under a single architecture. Kendall et al. [19] addressed the semantic segmentation problem through a multi-task deep learning model and assumed that the sub-tasks are independent and identically distributed, which successfully solved the problem of the precision being sensitive to the parameter weights. Lu et al. [4] also transformed the multi-scale scene learning problem into a multi-task learning problem for scene recognition. Multi-task learning frameworks have various advantages. It not only has the ability of implicit data augmentation and improved generalisation ability, but also helps the model to focus on the features that are relevant to the special traffic scene while excluding the unrelated ones, because auxiliary tasks can provide additional evidence of the relevance of features [16].

2.3 Dataset

In the autonomous driving field, the research relies heavily on some benchmarks (e.g. Kitti [10], Comma.ai [11], Cityscapes [12], DBNet [13] etc.). These datasets mainly concentrate on common problems in automatic driving tasks, such as pedestrian and vehicle detection, dynamic obstacle tracking, location and mapping etc. There are some datasets that are related to normal traffic scenes [14, 15]. For instance, the CamVid dataset [14] includes videos taken at daytime and at dusk in Cambridge, UK with a total of 701 tagged frames. However, to our best knowledge, there are few datasets that were built for special traffic scenes so far. Therefore, a new dataset for special traffic scenes is of necessity and was built up in our work.

3 Methodology

Behavioural decision-making is based on the outputs of the scene recognition task and plays an important role in an autonomous driving system. The behaviour of an autonomous vehicle (such as turn left, straight-forward, overtaking etc.) is inferred based on the single model has poor generalisation ability and is easy to have overfitting problems. To tackle this problem, three auxiliary tasks are adopted to improve the precision of the chief task, which are the time of occurrence (daytime or night-time), the weather type (dry, rainy, snowy or foggy) and the road attribute (rural, urban or highway).

Fig. 2 provides a parameter sharing network structure of the proposed method, which consists of three major steps. In the first step, given an input image, the corresponding feature map can be generated by a convolutional neural network (CNN). Each single task then shares the feature map followed by a task-specific sub-network. The final step merges multi-task branches by combining loss functions. In this section, we firstly introduce relevant notations and the network structure, followed by the loss function of each task. Moreover, a novel method to optimise the weight matching among different tasks is proposed. The training details and parameter adjustment methods will be given in Section 5.1.

3.1 Network architecture

Suppose shared feature maps are to be generated for T tasks and the corresponding scene category for the rth tasks is C_r. There are m training images \( X = \{x_1, ..., x_m\} \) with ground-truth labels \( Y = \{y_1, ..., y_m\} \) and \( y_i \subset\{1, ..., C\} \). That is, each scene image corresponds to T labels.

The multi-task learning network architecture is based on the parameter sharing framework [17], as shown in Fig. 2. As mentioned earlier, four tasks including one chief task and three auxiliary tasks are included, so each scene picture contains four information labels corresponding to four tasks, respectively. At first, multi-task sharing features are extracted through a pre-trained convolution network (e.g. VGG16, Resnet50). Subsequently, different from the traditional fully connected (FC) layer as a feature dimension reduction method, we use four separate global average pooling layers [21] to reduce the dimension of features extracted from the shared network and to generate four classification branches. In the end, the final classification result of each sub-network is squashed using Softmax function as the output of the multi-task network.

3.2 Multi-objective optimisation

Multi-task learning concerns the problem of optimising a model with respect to multiple objectives. The universal approach to combining multi-objective losses is to perform a weighted linear sum of the losses for each individual task:

\[
L_{\text{total}} = \sum_{t=1}^{T} \omega_t L_t
\]  

This is the frequently used strategy adopted by previous works [8, 22]. However, there are a number of issues with this approach. On the one hand, the linearcombination formulation is only sensible when there is a parameter set that is effective across all tasks. In other words, minimisation of a weighted sum of empirical risks is only valid if tasks are not competitive, which is rarely the case. Multi-task learning with conflicting objectives requires the modelling of the counterpoise between tasks, which is beyond what a linear combination actualizes [23]. On the other hand, the model performance is extremely sensitive to weight \(\omega_t\). To obtain the optimal values requires manually searching the hyperparameter space, which is very expensive. In the literature [19], homoscedastic uncertainty was used as a basis for weighting losses in a multi-task learning problem and the hyperparameter optimisation problem was successfully solved. Deng et al. [23] solved multi-task learning as multi-objective optimisation with the overall objective of finding a Pareto optimal solution, but this cannot be used for high dimensional data or for more task optimisation.

Following [19], we assume that the tasks are identically distributed and optimise the combined loss function by adding homoscedastic uncertainty in the proposed multi-task learning method. The softmax function is used to normalise the
classification results in the final output layer of each task. Finally, a cross-entropy loss is used as the cost function. The Softmax for this output can then be written as

\[
p(y = \hat{y}|g_\theta(x), \sigma) = \text{softmax}\left(\frac{1}{\sigma}g_\theta(x)\right)
\]

\[
= \frac{\exp\left(\frac{1}{\sigma}g_\theta(x)\right)}{\sum_c \exp\left(\frac{1}{\sigma}g_\theta(x)\right)}
\]  

(2)

The scalar \(\sigma\) balances the weights between multiple tasks and can be learnt. Using the cross-entropy function as loss function, we encode the label by one-hot, \(\hat{y}_i\) represents that the \(i\)th element of the output vector is 1, and the rest is 0. Based on the one-hot encoding method and the corresponding cross-entropy function, the final analytic formula for a single task is

\[
L(W, \sigma) = - y \log \hat{y}_i = - \log \hat{y}_i
\]

\[
= - \log \left(\frac{\exp\left(\frac{1}{\sigma}g_\theta(x)\right)}{\sum_c \exp\left(\frac{1}{\sigma}g_\theta(x)\right)}\right)
\]

\[
= - \frac{1}{\sigma} g_\theta(x) + \log \left(\sum_c \exp\left(\frac{1}{\sigma}g_\theta(x)\right)\right)\]

\[
= - \frac{1}{\sigma} g_\theta(x) - \log \sum_c \exp\left(g_\theta(x)\right)
\]

\[
- \frac{1}{\sigma} \log \sum_c \exp\left(g_\theta(x)\right) + \log \sum_c \exp\left(\frac{1}{\sigma}g_\theta(x)\right)
\]

\[
= \frac{1}{\sigma} L(W) + \log \sum_c \exp\left(g_\theta(x)\right)
\]

\[
= \frac{1}{\sigma} L(W) + \log \sigma
\]

(3)

where

\[L(W) = - \log \text{softmax}(y, g_\theta(x))\]

is for the cross-entropy loss. We adopt the simplified assumption according to [19]

\[
\frac{1}{\sigma} \log \sum_c \exp\left(\frac{1}{\sigma}g_\theta(x)\right) \approx \log \sum_c \exp\left(\frac{1}{\sigma}g_\theta(x)\right)^{1/\sigma^2}
\]

(4)

where \(W\) and \(\sigma\) are optimised at the same time. Finally, the maximum likelihood estimation of multi-tasks based on the cross-entropy loss function is

\[
L_{\text{total}}(W, \sigma_1, \ldots, \sigma_T)
\]

\[
= - \log(p(y_1, \ldots, y_T|g_\theta(x), \sigma_1, \ldots, \sigma_T))
\]

\[
= - \sum_{t=1}^T \log(p(y_t|g_\theta(x), \sigma_t))
\]

\[
= \sum_{t=1}^T \frac{1}{\sigma_t} L(W) + \sum_{t=1}^T \log \sigma_t
\]

(5)

4 Dataset

The dataset is intended for special traffic conditions and is largely different from previous ones for its novel hierarchy and favourable properties. This section briefly introduces the data acquisition method, the data grouping method and the structure of dataset.

Fig. 3 Number of images in the four tasks
(a) The time of occurrence, (b) The weather type, (c) The road attribute, (d) The special traffic scene

4.1 Data collection

Special traffic situations such as traffic accidents, road construction were not often encountered in normal driving environment. Therefore, collecting rich data for ideal visualisation is costly in terms of time, labour and hardware resources. On the other hand, since the occurrence of special traffic scene is unpredictable, it is difficult to collect sufficient data. For data collection, we adopted the method of combining online acquisition with real vehicle acquisition.

4.2 Structure of dataset

Fig. 3 demonstrates the details of the structure of the dataset collected exclusively for this paper. The dataset has 599 images for the four tasks and each task has several categories of driving conditions. All the images were taken for both daytime and night-time giving two categories for the auxiliary task of the time of occurrence (Fig. 3a). Weather type features four categories which are dry, rainy, snowy and foggy (Fig. 3b). Three road categories of urban road, rural road and highway are included in the attribute of road task (Fig. 3c). In Fig. 3d, the chief task special traffic scene has seven categories being normal road without any special traffic scene, highway toll station, ramp, regulation of traffic, traffic accident, road construction. At the same time, in this dataset, special traffic scenes with or without traffic signs each occupy 50% in order to test and improve the generalisation ability of the model.

5 Experiments and results

During our experiment, 80% of the data samples were used as the training set and the rest as the test set. For fair comparison with other methods, the same training and test sets were used for each comparative experiment.

5.1 Experiment and implementation

We demonstrated the efficacy of our method using the above dataset. During data pre-processing, all the images were firstly resized to 224×224×3 and then enhanced by random flipping, translating and scaling, followed by normalisation. For the network structure, the convolution layers of VGG16 and Resnet50 were used as shared networks. The convolution layer output is 7×7×512 for VGG16 and 7×7×2048 for Resnet50. Since the traditional FC layer needs enormous computing space, we adopted the average pooling layer [21] to generate 512 scalars and 2048 scalars for sub-task, respectively. The pre-trained model on the Imagenet dataset [24] was utilised as the initial value of the backbone network. The
model was tweaked with Adam optimiser [25] and the initial learning rate was $10^{-3}$.

Various methods exist for multi-task training, such as individual training, cross training and joint training. As the noise effect of different tasks is different, simply using the method of joint training cannot achieve the desired results. Hence, we first used the method of individual training to obtain the optimal loss (lowest average loss) for each task separately. Table 1 shows the optimal loss for each training task obtained with Resnet50 and VGG16 networks.

Normalising the reciprocal of the optimal loss values gives the ratio of each task as the initial value of $\sigma$ in formula (5). The method of joint training was next used with formula (5) as the loss function to automatically adjust the weights between the tasks.

The framework was built with Tensorflow and the experiment was carried out on a HP Z280 workstation with Xeon E5 CPU and KQR4000 (8 GB memory) GPU.

### 5.2 Result and comparison

In order to verify the effectiveness of the proposed method, it was compared with VGG + SVM, Resnet50 + SVM, BOW, VGG16 (single), Resnet50 (single) and BOW + kNN in terms of mean average precision (mPA). Feature extraction methods based on CNN and SIFT and discriminant models SVM and kNN were adopted. To ensure the fairness of comparative experiments, the training set and the test set were the same in all methods. The only difference lies in the fact that the proposed method had assistance from three auxiliary tasks in the training process, while the other methods only used the data in classification task for special traffic scene.

Table 2 shows the prediction precision of all the methods mentioned above for comparison. Since the size of the dataset used is limited, the overall accuracy of all methods is not very high with no >65%, but it serves the purpose of comparison. The accuracy of the method proposed in this paper is roughly 1 and 0.7% higher than that of VGG16 and Resnet50 trained with only a single task and is even more better than all the other methods. It is evident that the multi-task learning method still demonstrated higher accuracy than other methods.

The results showed that though the overall accuracy was relatively low due to the small size of the dataset, the multi-task learning method still further improved the performance under small-sized dataset.

### Acknowledgments

This work is supported by the National K&D Program of China under 2018YFB1305105 and the National Natural Science Foundation of China under Grants U1564214, 61751311, 61611540348.

### References

1. Zou, Q., Ling, H., Luo, S., et al.: ‘Robust nighttime vehicle detection by tracking and grouping headlights’, IEEE Trans. Intell. Transp. Syst., 2015, 16, (5), pp. 1–12
2. Di, S., Zhang, H., Li, C.-G., et al.: ‘Cross-domain traffic scene understanding: a dense correspondence-based transfer learning approach’, IEEE Trans. Intell. Transp. Syst., 2017, pp. 1–13
3. Lu, C., Lin, D., Jia, J., et al.: ‘Two-class weather classification’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2014
4. Lu, X., Li, X., Mou, L.: ‘Semi-supervised multitask learning for scene recognition’, IEEE Trans. Cybern., 2014, 45, (9), pp. 1967–1976
5. Zhu, Y., Liao, M., Yang, M., et al.: ‘Cascaded segmentation-detection networks for text-based traffic sign detection’, IEEE Trans. Intell. Transp. Syst., 2018, 19, (1), pp. 209–219
6. Luo, H., Yang, Y., Tong, B., et al.: ‘Traffic sign recognition using a multi-task convolutional neural network’, IEEE Trans. Intell. Transp. Syst., 2018, 19, (4), pp. 1100–1111
7. Arcos-Garcia, A., Alvarez-Garcia, J.A., Sorta-Morillo, L.M.: ‘Deep neural network for traffic sign recognition systems: an analysis of spatial transformers and stochastic optimisation methods’, Neural Netw., 2018, 99, pp. 158–165, S0893608018300054
8. Liu, Y., Kodagoda, S., Wang, Y., et al.: ‘Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks’. 2016 IEEE Int. Conf. on Robotics and Automation (ICRA), 2016
9. Lee, Y., Kim, G.: ‘Fog level estimation using non-parametric intensity curves in road environments’, Electron. Lett., 2017, 53, (21), pp. 1494–1496
10. Geiger, A., Lenz, P., Urtasun, R.: ‘Are we ready for autonomous driving?’ The KITTI vision benchmark suite’. 2012 IEEE Conf. on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 3354–3361
11. Santanaand, E., Hotz, G.: ‘Learning a driving simulator’. CoRR, doi: abs/1608.01230, 2016
12. Cordts, M., Omran, M., Ramos, S., et al.: ‘The cityscapes dataset for semantic urban scene understanding’. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 3213–3223

Table 1 Optimal loss for each task obtained individually with VGG16 and Resnet50

| Shared networks | Task1 | Task2 | Task3 | Task4 |
|-----------------|-------|-------|-------|-------|
| VGG16           | 0.32  | 1.25  | 2.54  | 1.67  |
| Resnet50        | 0.41  | 2.02  | 2.34  | 3.26  |

Table 2 Comparison of mean average precision

| Method            | mAP, % |
|-------------------|--------|
| BOW + SVM         | 51.2   |
| BOW + kNN         | 50.5   |
| VGG16 + SVM       | 59.3   |
| Resnet50 + SVM    | 60.2   |
| VGG16 (single)    | 60.3   |
| Resnet50 (single) | 61.7   |
| ours              | 62.3   |
| VGG16 (multi)     | 63.4   |
| Resnet50 (multi)  |        |
[13] Chen, Y., Wang, J., Li, J., et al.: ‘LiDAR-video driving dataset: learning driving policies effectively’. 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 5870–5878

[14] Brostow, G.J., Fauqueur, J., Cipolla, R.: ‘Semantic object classes in video: a high-definition ground truth database’, Pattern Recognit. Lett., 2009, 30, (2), pp. 88–97

[15] Di, S., Zhang, H., Mei, X., et al.: ‘A benchmark for cross-weather traffic scene understanding’. 2016 IEEE 19th Int. Conf. on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 2016, pp. 2150–2156

[16] Baxter, J.: ‘A model of inductive bias learning’, J. Artif. Intell. Res., 2000, 12, pp. 149–198

[17] Caruana, R.A.: ‘Multitask learning: a knowledge-based source of inductive bias’, Mach. Learning Proc., 1993, 10, (1), pp. 41–48

[18] Kokkinos, I.: ‘Ubernet: training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory’. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2017

[19] Kendall, A., Gal, Y., Cipolla, R.: ‘Multi-task learning using uncertainty to weigh losses for scene geometry and semantics’. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2018

[20] Xu, X., Zuo, L., Li, X., et al.: ‘A reinforcement learning approach to autonomous decision making of intelligent vehicles on highways’, IEEE Trans. Syst., Man, Cybern.: Syst., 2018, PP, (99), pp. 1–14

[21] Lin, M., Chen, Q., Yan, S.: ‘Network in network’. arXiv preprint arXiv:1312.4400, 2013

[22] Teichmann, M., Weber, M., Zoellner, M., et al.: ‘Multinet: real’. 2018 IEEE Intelligent Vehicles Symp. (IV) IEEE, 2018

[23] Sener, O., Kolotou, V.: ‘Multi-task learning as multi-objective optimization’, Adv. Neural. Inf. Process. Syst., 2018

[24] Deng, J., Dong, W., Socher, R., et al.: ‘Imagenet: a large-scale hierarchical image database’. IEEE Conf. on Computer Vision & Pattern Recognition, 2009

[25] Kingma, D.P., Ba, J.: ‘Adam: a method for stochastic optimization’, Comput. Sci., 2014

J. Eng., 2020, Vol. 2020 Iss. 13, pp. 471-475
This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)