Signal Processing 147 (2018) 146–153

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Vehicle color recognition using Multiple-Layer Feature Representations of lightweight convolutional neural network

Qiang Zhang[a,c], Li Zhuo[a,b,c,*], Jiafeng Li[a,c], Jing Zhang[a,c], Hui Zhang[a,c], Xiaoguang Li[a,c]

[a] Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing, China
[b] Collaborative Innovation Center of Electric Vehicles in Beijing, Beijing, China
[c] College of Microelectronics, Faculty of Information Technology, Beijing University of Technology, Beijing, China

ARTICLE INFO

Article history:
Received 30 July 2017
Revised 6 December 2017
Accepted 16 January 2018

Keywords:
Multiple-Layer Feature Representations
Lightweight convolutional neural network
Vehicle color recognition
Spatial Pyramid Matching

ABSTRACT

In this paper, a vehicle color recognition method using lightweight convolutional neural network (CNN) is proposed. Firstly, a lightweight CNN network architecture is specifically designed for the recognition task, which contains five layers, i.e. three convolutional layers, a global pooling layer and a fully connected layer. Different from the existing CNN based methods that only use the features output from the final layer for recognition, in this paper, the feature maps of intermediate convolutional layers are all applied for recognition based on the fact that these convolutional features can provide hierarchical representations of the images. Spatial Pyramid Matching (SPM) strategy is adopted to divide the feature map, and each SPM sub-region is encoded to generate a feature representation vector. These feature representation vectors of convolutional layers and the output feature vector of the global pooling layer are normalized and cascaded as a whole feature vector, which is finally utilized to train Support Vector Machine classifier to obtain the recognition model. The experimental results show that, compared with the state-of-art methods, the proposed method can obtain more than 0.7% higher recognition accuracy, up to 95.41%, while the dimensionality of the feature vector is only 18% and the memory footprint is only 0.5%.

1. Introduction

Vehicle recognition plays an important part in the applications of Intelligent Transportation System, criminal investigation and so on. Its task is to recognize the type, color and license plate of the targeted vehicle. Color is one of the basic attributes of vehicles, therefore, color recognition plays a significant role in vehicle recognition. However, due to complex impacts of illuminations, weather conditions, noises and image capture qualities, vehicle color recognition has become a challenging task. Firstly, illumination, noises and special weather conditions will lead to obvious changes in visual appearance of vehicle colors. For instance, the vehicle colors under different illuminations vary greatly, which makes the vehicle recognition very difficult. Secondly, considering the differences in camera positions and parameters, there are significant differences in angles and focal distances. In addition, the size and area of vehicles in the image change greatly, which poses another great challenge to vehicle color recognition. Finally, the interference of complex scenes and non-vehicle color components further escalate the difficulty of vehicle color recognition.

In recent years, many researchers have proposed various solutions for vehicle color recognition. In general, the research on vehicle color recognition can be divided into two stages:

The first stage is based on handcrafted features combined with classifiers [1,4–6], which focus on designing various features of vehicle colors manually, such as color histogram [1], color moment [2] and color correlogram [3], and then these features are used to train the classifiers, such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Artificial Neural Network (ANN). Chen et al. [1] extracted the color histogram features and introduced spatial information using Spatial Pyramid Matching (SPM) and Feature Context (FC) to establish Bag of Words (BoW) model, then combined linear SVM to solve the vehicle color recognition problem. Baek et al. [4] also extracted the color histogram of H and S components in HSV color space to construct the bi-dimensional feature vectors and further combined SVM classifiers for vehicle color recognition. Another approach was proposed by Dule et al. [5], which extracted different color features from various color spaces, and tested the vehicle color recognition performance using different classifiers, including KNN, ANN and SVM. The experimental result indicates that ANN classifier combined with 8-bin color

* Corresponding author.
E-mail addresses: zhuoli@bjut.edu.cn (L. Zhuo), zhj@bjut.edu.cn (J. Zhang), huizhang@bjut.edu.cn (H. Zhang), lxg@bjut.edu.cn (X. Li).

https://doi.org/10.1016/j.sigpro.2018.01.021
0165-1684/

histogram feature vector can obtain the highest recognition accuracy.

Apart from above vehicle color recognition method based on global feature extraction, Hu et al. [6] provided an alternate way of vehicle color recognition by constructing the color reflection model to remove the non-vehicle color areas in the image, such as background, vehicle wheels and windows, then directly extracting features from the major vehicle color regions to train the SVM classifier.

These methods mentioned above basically adopt handcrafted features. Therefore, they are tending to achieve higher execution speed but weaker generalization capability and lower recognition accuracy. In addition, the design of handcrafted features requires professional knowledge. An appropriate feature usually takes massive experience and time to validate its effectiveness for a certain task. Therefore, it is difficult to manually design proper features for new data and tasks.

The second stage is based on deep learning. The research of this stage mainly focuses on two aspects: the first is to make use of deep neural network to obtain the feature representation of the images then apply it for vehicle color recognition in combination of traditional classifier. Hu et al. [7] proposed a vehicle color recognition method, in which the deep features of AlexNet [8], and kernel SVM are used, combined with SPM learning strategy. The other is to use an end-to-end deep neural network structure, which merges the feature extraction and classifier into a unified framework through joint optimization. Rachmadi et al. [9] designed a parallel CNN network to achieve end-to-end vehicle color recognition, which learns the recognition model from big data by using two convolutional networks, and integrated the two parts by a fully connected layer. The research results indicate that for vehicle color recognition, compared to the traditional recognition methods of handcrafted features + classifier, the feature representations learnt from deep learning have strong generalization capability and the recognition performance can be improved obviously.

The representative deep network structures used in vehicle color recognition mainly include AlexNet [8], GoogleNet [13], VGG-Net [14], and so on, which are originally designed for complex classification tasks and characteristic of a massive amount of data. Therefore, the network structures usually possess large number of parameters and are easily subject to the occurrence of over-fitting phenomenon. In addition, they require large computational and storage resources. But, recently, with the advance of research, some researchers found that for some specific tasks or applications, a lightweight network structure can also achieve the most advanced result [10].

Deep neural network demonstrates strong learning ability and highly efficient feature extraction capability, which can extract information from low-level raw data to high-level abstract semantic concepts. The hierarchical feature representation can entitle it with prominent advantages when extracting global features and context semantic information of the images. However, the existing methods commonly make use of the output features of the last layer for recognition while neglecting the feature information of the previous layers. Actually, these features of lower layers contain considerable information of images, which may promote the recognition performance. However, if all the features from the intermediate layers can be employed, the extremely high dimension of feature vectors will result in training the classification model too difficult or even failure. To address this problem, a traded-off solution is proposed in this paper. Firstly, a lightweight CNN network architecture is designed for the vehicle color recognition task, which contains five layers, i.e. three convolutional layers, a global pooling layer and a fully connected layer. The feature maps of convolutional layers are used for feature representation of the image. Specifically, SPM strategy is adopted to divide the feature maps of the convolutional layers into four sub-regions. Then, each sub-region is encoded to obtain a feature representation vector of the layer. Next, these feature representation vectors and the output feature vector of the final global pooling layer are normalized and cascaded as a whole feature vector to represent the content of the image. Finally, the linear SVM is used as the classifier for vehicle color recognition. The experimental results demonstrate that, the features from the intermediate layers can help to improve the recognition accuracy. And compared with the state-of-art methods, the proposed method can get higher recognition accuracy by more than 0.7%, up to 95.41%, while with the lower dimensionality of the feature vector and smaller memory size of CNN model.

Table 1
Comparison results of the proposed network structure and AlexNet structure.

	AlexNet	Proposed network
Input	227 × 227 × 3	227 × 227 × 3
Layer 1	conv,96	conv, 48
Layer 2	conv,256	conv, 128
Layer 3	conv,384	conv, 192
Layer 4	conv,384	–
Layer 5	conv,256	–
Layer 6	fc,4096	GAP,192
Layer 7	fc,4096	–
Output	fc,8	fc,8
Memory	227.6M	1.1M

The rest of this paper is organized as follows. Section 2 describes the main ideas and details of the proposed method. Experimental results and analysis are presented in Section 3. Finally, the conclusions are drawn in Section 4.

2. The proposed method

Convolutional neural network has become the dominant machine learning approach for visual recognition tasks. To fulfill complex tasks, which usually need to identify hundreds or even thousands of categories, CNN model usually has a large number of parameters. When used for small and medium size datasets, overfitting often occurs. In vehicle color recognition, color categories and the size of the datasets are both limited. For example, in literature [6], vehicle color contains red, yellow, blue and green, totally four categories. And in [5], the vehicle colors contains white, black, gray, red, blue, green, and yellow, totally seven categories. In this paper, the public vehicle color dataset in [1,7,9] is adopted, which includes eight color categories to recognize. In order to achieve a good tradeoff between the performance and the computational complexity, a lightweight CNN network architecture is designed to extract the features for vehicle color recognition, which includes three convolutional layers, a global pooling layer and a fully connected layer, totally five layers.

The framework of the proposed method is depicted in Fig. 1. Considering that the feature maps of the convolutional layers contains rich information of the vehicle images, all feature representations of the convolutional layers are employed and combined with the output feature vector of the global pooling layer to form a whole vector to represent the content of the images. The linear SVM classifier is used to train the classification model. In the following, the details of the proposed method will be described.

2.1. Lightweight convolutional neural network structure design

The lightweight convolutional neural network architecture designed in this paper is shown with the black dotted line in Fig. 1. The number of neurons at the three convolutional layers is 48, 128,

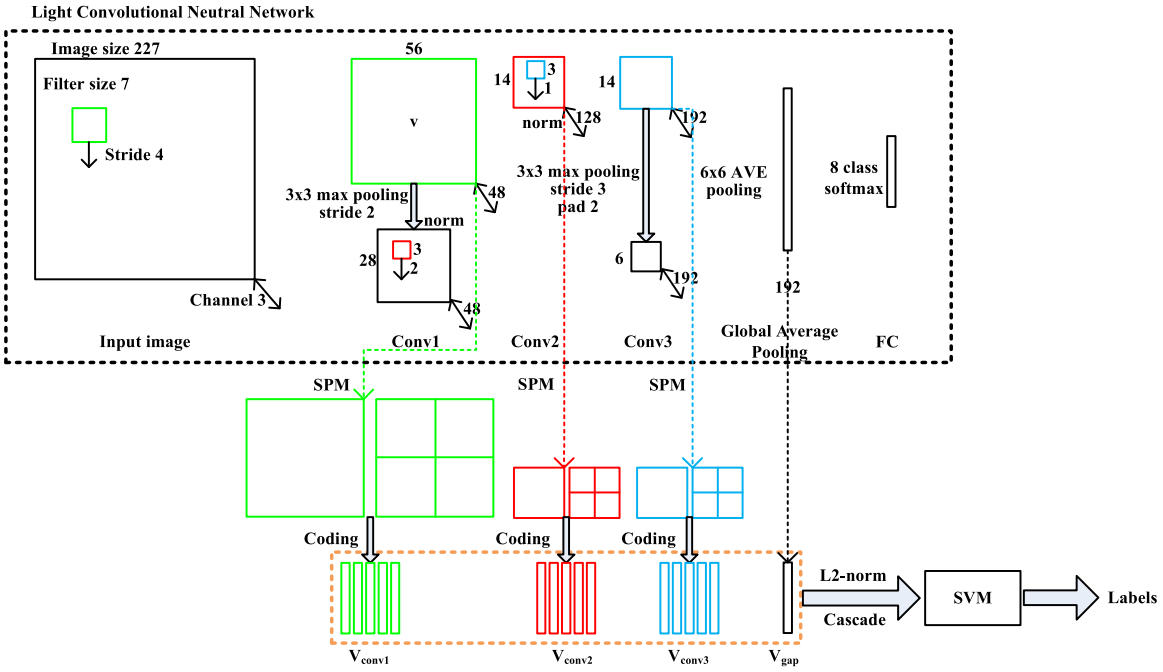


Fig. 1. Framework of the proposed vehicle color recognition method.

and 192 respectively. Table 1 shows the comparison results of the proposed network with the classical AlexNet [8] network structure.

It can be seen from the Table 1 that, the proposed lightweight CNN network structure is mainly illustrated in three aspects. Firstly, this structure reduces the number of convolutional layers and corresponding active layers, which requires fewer non-linear operations. This can decrease the computational complexity and improve the learning speed of the network to some degree. Meanwhile, it lessens the number of convolutional kernels in the convolutional layers, which declines the amount of the optimized parameters in learning process, thus increasing the learning speed of the network and reducing the risk of over-fitting. Finally, the lightweight CNN network adopts a global average pooling layer as the link layer between the third convolutional layer and the final fully connected layer, which can avoid concentrating the majority of network parameters onto the fully connected layer. By these ways, the network structure can be greatly simplified. The memory footprint by the lightweight network proposed in this paper is only 1.1 M, far smaller than the 227.6 M of AlexNet. How to design each component of lightweight CNN network structure is detailed as follows.

2.1.1. Convolutional layer

As an indispensable part of CNN, convolutional layer serves to enhance the original features of signals and reduce the noises. Multiple image features can be extracted from diverse convolutional kernels. For the proposed CNN network structure in this paper, the convolutional layer performs discrete convolutional operation on the input signal. As for an input image I, if K is the convolutional kernel at the current layer, the output image H obtained after the discrete convolutional operation can be expressed as:

$$\mathrm{H}[x,y]=\sum_{m=0}^{w-1}\sum_{n=0}^{h-1}\mathrm{I}[m,n]\mathrm{K}[x-m,y-n] \tag{1}$$

where x and y are the coordinates of input image, m and n are the coordinates of convolutional kernel respectively. Updating and iterations obtain the kernel parameters K during the training process. In addition, since the hierarchical non-linear operation of CNN is implemented by the activation function, the selection of the activation function poses huge influence on the final recognition performance. The common activation functions in deep learning include Sigmoid, tanh and ReLU (Rectified Linear Units). In this paper, ReLU function is adopted as the activation function to activate the neurons.

2.1.2. Normalized layer

The network model parameters are updated constantly during the training process, which often leads to shift in input data distribution of each subsequent layer. Meanwhile, the learning process requires the operation on each layer to comply with input data distribution. Therefore, to overcome the influence of input data distribution on recognition performance, the data are normalized prior to the convolutional layer respectively in this paper. Firstly, the average of images is subtracted from the input image; then, before the convolutional operation on the second and third layers, the Local Response Normalization operation (LRN) is performed on the input data. The normalization operation can make the data distribution more rationally and more easily to be distinguished. Moreover, the network convergence rate and recognition accuracy can be improved during the training process. Denote $a^i_{x,y}$ as the activity of a neuron computed by applying kernel i at position (x, y), and $b^i_{x,y}$ the normalization result is given by the Eq. (2):

$$b^i_{x,y}=a^i_{x,y}/\left(k+\alpha\sum_{j=max(0,i-n/2)}^{min(N-1,i+n/2)}\left(a^j_{x,y}\right)^2\right)^{\beta} \tag{2}$$

where the sum that runs over n "adjacent" kernel maps at the same spatial position, and N is the total number of kernels in the layer. The constants k, n, α, and β are hyper-parameters whose values are determined using a validation set. In this paper, they are set as $k = 2$, $n = 5$, $\alpha = 10^{-4}$, and $\beta = 0.75$ respectively.

2.1.3. Pooling layer

The common pooling operations include average pooling, maximum pooling and random pooling, etc.. In the proposed CNN network structure, maximum pooling operation is carried out on the

output feature maps of the first and third convolutional layers. The pooling operation can reduce the dimensionality of the feature map, accelerate convergence rate, lower down computational complexity and provide certain rotation invariance.

2.1.4. Fully connected layer and loss function

The convolutional layer, pooling layer and activation function serves to map the raw data onto the feature space in the hidden layer while the fully connected layer maps the "distributed feature representation" learned onto the marked sample space. In other words, the label information of the categories from the target loss function will be transmitted to the previous convolutional layers through fully connected layer. Therefore, a fully connected layer is usually set up at the last layer of CNN to realize mapping between label information and the feature space.

On the other hand, there are generally multiple fully connected layers with too many parameters in traditional CNN network, easily leading to over-fitting (for instance, AlexNet contains three fully connected layers). To reduce the number of parameters of the fully connected layer, in this paper, there designs a global average pooling (GAP) layer [11] between the third convolutional layer and the final fully connected layer. Global average pooling operation is carried out on the feature map of the third convolutional layer so that each feature map can output only an average value, equivalent to a dimensionality reduction operation, which can fasten the parameter learning process of the fully connected layer, and thus, can sharply decrease the quantity of network parameters and avoid the risk of over-fitting.

Fully connected layer serves as a classifier in CNN while the selection of loss function directly determines the performance of classification model learned by the network. The common loss functions include Euclidean loss used in real value regression, Triplet loss used in human facial identification and Softmax loss used in single-label recognition. Considering that vehicle color recognition belongs to single label sample categorization, in this paper Softmax function is adopted as the loss function:

$$p\left(y^{(i)}=j|x^{(i)};\theta\right)=e^{\theta_j^T x^{(i)}}/\sum_{l=1}^{k}e^{\theta_l^T x^{(i)}} \tag{3}$$

where p is the probability of input sample $x^{(i)}$ belonging to jth category, and θ is the parameter of the network layer.

Overall, our proposed CNN structure contains 5 layers. The first layer uses $7\times7\times3$ kernel with total 48 kernels, second layer and third layer use $3\times3\times48$ kernel with total 128 kernels, $3\times3\times128$ kernel with total 192 kernels, respectively. Pooling operations after first convolutional layer use size of 3×3 with 2 pixel strides and after third convolutional layer 3×3 with 3 pixel strides respectively. At global average pooling layer, the average pooling kernel size is set as $6\times6\times192$. And the number of neurons of fully connected layer is 8, with the same size of vehicle color categories to be recognized. The input of the network is a 3-channel color image whose resolution is $227\times227\times3$. In our method, the proposed CNN architecture is trained using the examples to extract the features. For each input image, the output of each functional layer, such as convolutional layers and pooling layers, are regarded as the features and used for vehicle color recognition. The size and dimension of the features at different layer are shown in Table 2.

2.1.5. Visualization of convolutional Kernels and feature maps

To better understand how the proposed CNN network extracts the color information, the convolutional kernels of the first convolutional layer are visualized in Fig. 2. As seen in Fig. 2, the first convolutional layer can extract rich color features of the input image. All vehicle color variations in dataset are presented in the kernels. In other words, the proposed lightweight CNN structure can effectively extract rich color feature information for vehicle color recognition.

Table 2
The size and dimension of the features at different layers.

Layer	Kernel	Stride	Feature Map Size	Dimension
Conv1	7×7	4	$56\times56\times48$	150,528
Pooling 1	3×3	2	$28\times28\times48$	37,632
Conv2	3×3	2	$14\times14\times128$	25,088
Conv3	3×3	1	$14\times14\times192$	37,632
Pooling 2	3×3	3	$6\times6\times192$	6912
GAP	6×6	1	$1\times1\times192$	192

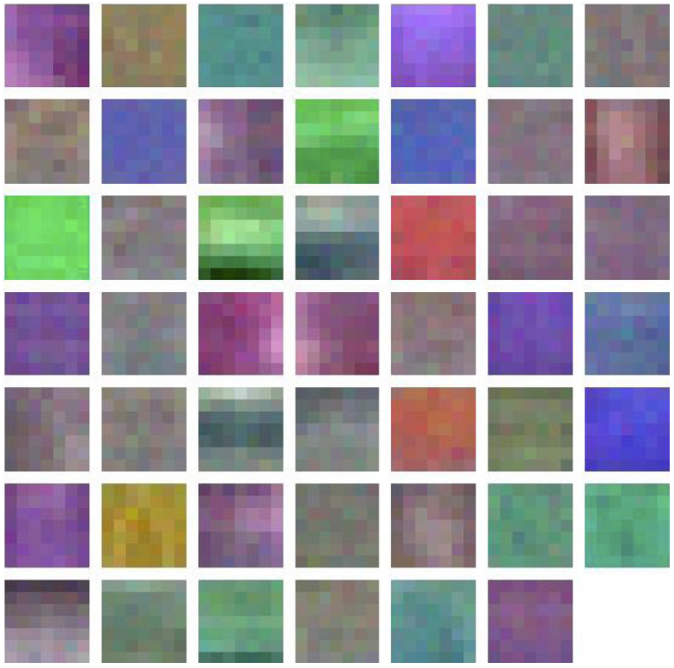


Fig. 2. Visualization of 48 convolutional kernels of the first convolutional layer. The convolutional kernels of size $7\times7\times3$ are learned on the $227\times227\times3$ input images.

One of the main challenges to vehicle color recognition is to reduce the interference of non-vehicle color regions in the image, such as background, vehicle windows and wheels. In the traditional methods based on handcrafted features, the interference reduction methods mainly include two categories: the first is to divide the image into sub-regions and determine the weights of each sub-region by training classifiers in order to reduce the weights of non-vehicle color regions [1]; the second is to construct a mathematical model to measure the correlation between the vehicle color and non-vehicle color regions, then, removing the non-vehicle color regions [6] according to the measurement results.

To verify the feature representation capability proposed in this paper, the feature maps of the convolutional layers are visualized. Firstly, the average feature map of each convolutional layer is calculated according to Eq. (4), and then visualized by heat-map. The visualization result is shown in Fig. 4, where three average convolutional feature maps are listed.

$$F_i=\frac{1}{c^i}\sum_{l=1}^{c^i}f_i^j \tag{4}$$

where f_i^j is the jth feature map of ith convolutional layer, c^i is the number of convolutional kernels of ith layer, F_i is the average feature map of ith layer and its size is consistent with that of a single feature map at ith layer. For instance, the size of feature maps at

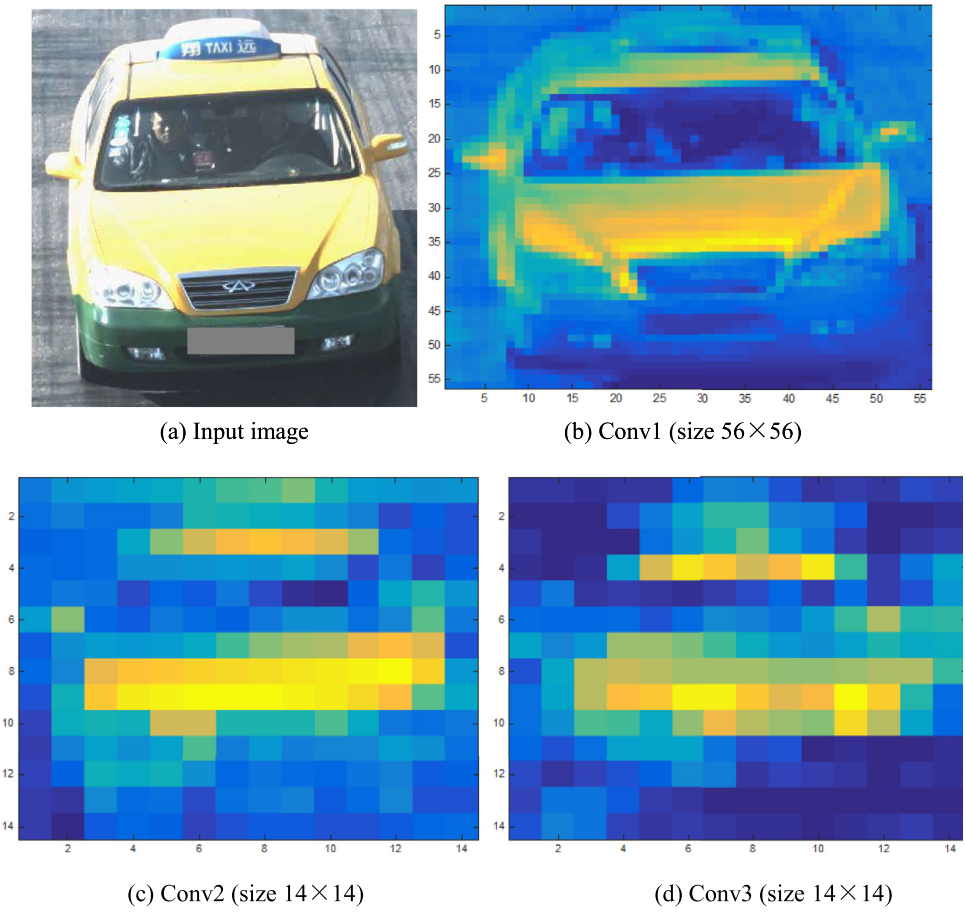


Fig. 3. Visualization of three convolutional average features.

the third convolutional layer is $14 \times 14 \times 192$ and then the size of average feature map is 14×14, as shown in Fig. 3(d).

It can be found in Fig. 3 that, larger feature value in heat-map (yellow region in Fig. b, c and d) corresponds to the main vehicle color regions while smaller feature value (dark region) corresponds to non-vehicle color regions, such as background and windows, and distorted color region caused by illumination. This indicates that the color features extracted are robust to the interference.

The average feature map of the second pooling layer is visualized in Fig. 4 and the heat-map value occupied by each point is marked as shown. It can be seen that the features obtained from the proposed CNN network can determines the corresponding weights adaptively according to the vehicle color contribution in different regions, which means the excellent representation capability.

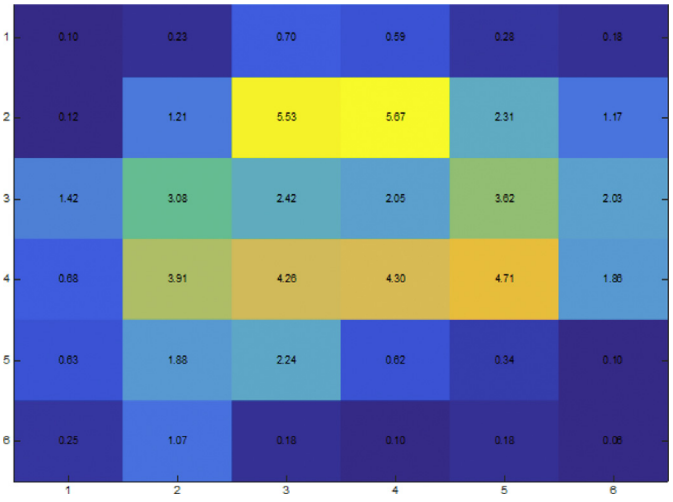


Fig. 4. Visualization of the average feature map of the second pooling layer.

2.2. Convolutional layer feature representation based on SPM

As described above, deep neural network has strong learning capability and highly efficient feature representation capability, which can extract information from low to high level. If all the features from the intermediate layers can be utilized, the feature representation capability can be enhanced to a certain extent, but it also results in high dimension of feature vectors. Therefore, the existing methods commonly adopt the features output from the last layer of deep network for recognition, but neglect the features output from the intermediate layers. To make full use of these features, this paper proposes a feature representation method for the convolutional layer based on SPM strategy. SPM is a classical method that embeds spatial information into feature vector [12]. It aims to extract features from different sub-regions and aggregating the features of all the regions together to describe an image. In order to reduce the dimensionality of the features, the proposed method divides the feature map into four sub-regions using SPM, and then encodes the sub-regions to obtain the feature representation vector of the convolutional layer.

Table 3
The number of images for each color.

Color	Black	Blue	Cyan	Gray	Green	Red	White	Yellow	Total
Number	3442	1086	281	3046	482	1941	4742	581	15,601

The proposed method divides the CNN convolutional feature map (as shown in Fig. 1) using SPM, whose advantage is that only one deep feature extraction process is required, avoiding multiple feature extractions. At the same time, the spatial information is embedded into the representation vector, thus, improving the description and discrimination capability.

In this paper, we use the Eq. (5) to encode each SPM division and obtain a compact feature representation of the feature map:

$$P_i = \frac{f_i^k}{w \times h}, \quad k \in 1, 2, \ldots, c^i \tag{5}$$

where f_i^k is the kth feature map of the ith layer, $w \times h$ is the size of the current feature map, and c^i is the number of neurons of the ith layer convolutional layer. P_i represents the output of encoding and has dimensions $1 \times 1 \times c^i$. For example, the feature representation vector of the third convolutional layer is a $1 \times 1 \times 192$ dimensions vector.

After L2-normalization, the feature representation vectors of each convolutional layer and the output feature vector of the global pooling layer are cascaded as a whole vector to represent the content of the image. L2 normalization operation can be expressed as:

$$x'_j = x_j / \sqrt{\sum_{j=1}^{c^i} x_j^2} \tag{6}$$

where x_j is the jth eigenvalues of P_i, and x'_j represents the values after the L2 normalization operation.

We use v_{GAP} to represent the normalized feature value of the global pooling layer. Similarly, $v_{conv1}, v_{conv2}, v_{conv3}$ represent the different normalized feature of the convolutional layers. The feature cascade process can be expressed as:

$$V = [v_{conv1},\ v_{conv2},\ v_{conv3},\ v_{GAP}] \tag{7}$$

where the dimensions of $v_{conv1}, v_{conv2}, v_{conv3}$ and v_{GAP} are 240(48×5), 640(128×5), 960(192×5), and 192 respectively. Thus, the dimension of the combined feature V is 2032. Finally, the linear SVM is trained as a classifier using the combined feature vector.

3. Experimental results and analysis

To verify the effectiveness of the vehicle color recognition method proposed in this paper, we conducted the experiments on a public Vehicle Color dataset [1] provided by Chen et al. The experimental results are compared with the state-of-the-art vehicle color recognition methods. The comparative experimental platform is set as following: 3.3-GHZ 4-core CPU, 8GB-RAM, Tesla-K20C GPU, and Ubuntu 64-bit operating system. Next, each experiment will be introduced in details.

3.1. Vehicle color dataset

In this paper, a public Vehicle Color dataset is adopted for experimental verification and comparison [1,7]. This dataset contains 15,601 vehicle images, including black, blue, blue green, gray, green, red, white and yellow, totally eight color categories. Table 3 shows the number of images for each color, in which there are 282 cyan vehicle images with the minimum proportion, and 4743 white vehicles with the maximum proportion. All the images in the dataset come from front images (or slight angle change) captured from road monitoring system. Besides, each image only contains one vehicle. The dataset has lots of environmental changes, such as weather condition and illumination. Some examples of the Vehicle Color dataset are shown in Fig. 5. There are multiple vehicle types in the dataset, such as trucks, cars and buses, which pose great challenge to vehicle color recognition. In the experiments, to compare and study the proposed method with the other methods, the same setting is adopted, that is to divide the dataset into training data and testing data with ration of 1:1 [1,7,9].

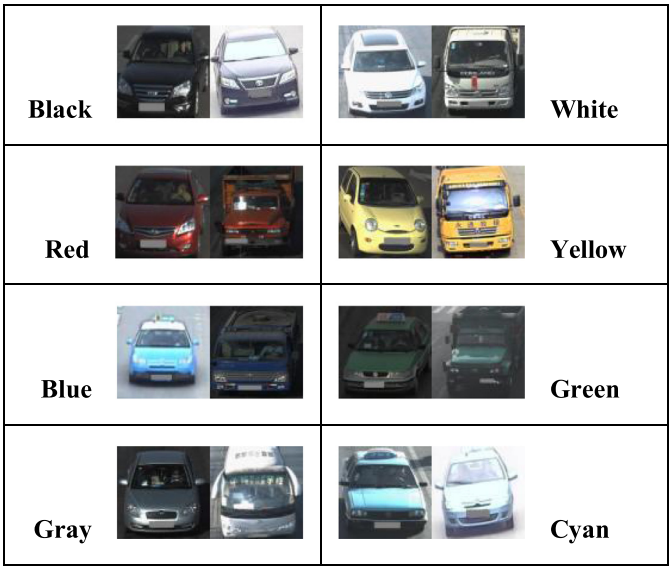


Fig. 5. Some examples of the Vehicle Color dataset.

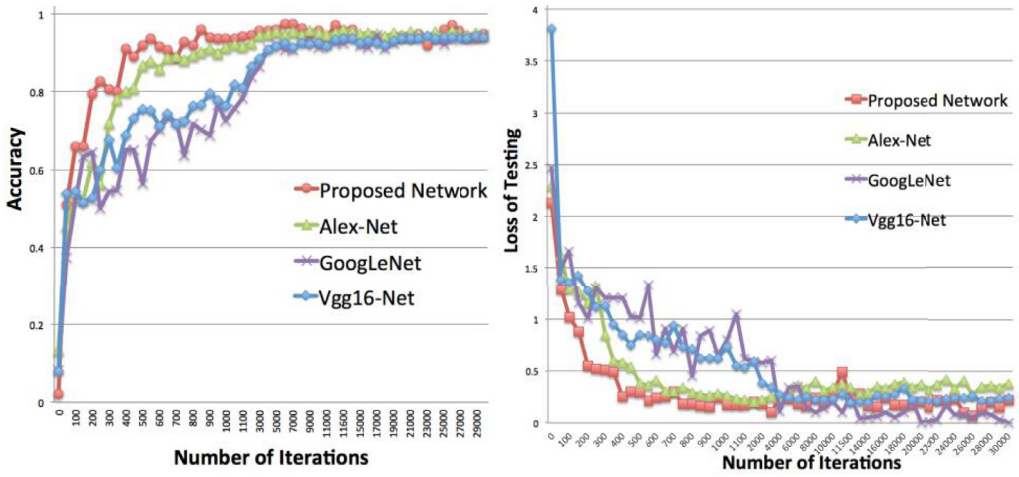
3.2. Performance of lightweight convolutional neural network

To verify the recognition performance of lightweight convolutional neural network proposed in this paper, a comparison experiment is carried out on this network and several other commonly used CNN networks, including AlexNet [8], GoogleNet [13] and VGG-Net [14], whose depth is 8, 22 and 16 respectively. The CNN networks are all constructed on Caffe (Convolutional Architecture for Fast Feature Embedding) platform [15]. In the training process, the same random initialization parameters and the iteration training network parameters of random gradient lowering algorithm are adopted. The trained model is used for vehicle color recognition. Fig. 6 shows the experimental comparison results. It should be noted that, for fair comparison, the CNN networks of AlexNet, GoogLeNet and Vgg16-Net all don't use pre-training strategy, but only end-to-end training mode.

It can be seen from Fig. 6(a) that, compared to the existing deep CNN networks, the proposed lightweight CNN network can even obtain higher recognition accuracy, reaching 94.73%, while the depth is only 5, much shallower and lighter than the other three CNN networks. In addition, it can be seen from Fig. 6(b) that, the proposed network is characteristic of faster network convergence rate and shorter computational time during vehicle color recognition. This is because the proposed lightweight CNN network has simpler structure, fewer parameters and much lower complexity.

3.3. Impacts of deep features at different layers on recognition results

To verify the impact of different features on vehicle color recognition accuracy, the framework of "deep features + SVM" is adopted



(a) Comparison results of the recognition accuracy (b) Comparison results of the testing loss

Fig. 6. Comparison results of different networks for vehicle color recognition.

Table 4
Comparison of the recognition accuracy using different deep features.

Layer	Black	Blue	Cyan	Gray	Green	Red	White	Yellow	AP
Conv1	0.9402	0.8656	0.8500	0.7242	0.7635	0.9763	0.8996	0.9241	0.8679
Pooling 1	0.9518	0.9116	0.9214	0.7597	0.7884	0.9711	0.8979	0.9103	0.8890
Conv2	0.9582	0.930	0.9643	0.8398	0.8340	0.9845	0.9350	0.9690	0.9269
Conv3	0.9698	0.9687	1	0.8411	0.7925	0.9887	0.9494	0.9655	0.9345
Pooling 2	0.9669	0.9761	0.9929	0.8575	0.8257	0.9918	0.9464	0.9689	0.9408
GAP	0.9797	0.9705	0.9929	0.8582	0.8423	0.9918	0.9688	0.9793	0.9479
Multiple Layers	0.9756	0.9834	0.9857	0.8884	0.8672	0.9938	0.9629	0.9759	0.9541

in this paper, which is mainly consisted of deep feature extraction and SVM classifier. The proposed CNN architecture is used to extract the deep features, and the SPM strategy and encoding processing are used to obtain a feature representation vector of each layer. Specifically, all feature representations of the convolutional layers are employed and combined with the output feature vector of the global pooling layer to form a whole vector to represent the content of the images. The recognition accuracy using the features from different layers are shown in Table 4.

It can be seen from Table 4 that, the less the layer, the weaker the feature representation and distinction capability and the lower the recognition accuracy. Compared with the method only using the features of a final single layer, the combined features from multiple layers can effectively improve the recognition accuracy, up to 0.62%. This is because the features from the intermediate layers can provide other supplemental information to the final global features, thus, better recognition performance can be obtained.

The optimal recognition performance in the table using the features of a single layer is obtained when applying the output features from global average pooling layer to train SVM classifier. This result indicates that global average pooling operation can not only decrease the network complexity and reduce the dimension of features but also effectively maintain the feature representation capability.

3.4. Comparison of recognition performance of different methods

To verify the effectiveness of the proposed method in this paper, we performed the comparison experiment with the state-of-the-art vehicle color recognition methods, including:

(1) The method proposed in [1], where the color histogram is combined with Feature Context strategy to construct Bag of Words model. The vehicle color recognition is performed based on linear SVM and the 12,288-dimensional feature vector.
(2) The method proposed in [9], where a parallel CNN network is proposed to achieve end-to-end vehicle color recognition. Its parallel structure is embodied in adopting two convolutional networks for learning. The two channels of output features are inputted into a fully connected layer for data integration and totally the dimension of feature vector is 4096.
(3) The method proposed in [7], which combines deep features and Kernel-SVM to train the classification model. SPM strategy is proposed to improve the representation capability of the feature vector. The dimension of feature vector is 11,520.

In our proposed method, linear SVM is chosen as the classifier. And the dimension of feature vector is 2032. Table 5 shows the experimental comparison result on Vehicle Color dataset using the four methods.

It can be seen from Table 5 that, compared with other three methods, the proposed method can achieve the highest recognition accuracy with the lowest feature dimension. Especially, when compared with the state-of-art method in [7], the proposed method can obtain even more than 0.7% better recognition accuracy, while the dimension of the features is only 18% and the memory footprint by proposed CNN network is only 0.5%.

Even more, from Tables 4 and 5, it also can be seen that, when only using the features of GAP layer for color recognition, it still can achieve a lightly higher recognition accuracy than the method in [7], while the feature dimension of GAP layer is only 192, far lower than the 11,520-dimension deep feature adopted by the comparison method. It can be concluded that the designed lightweight CNN network can efficiently extract the features of the vehicle colors.

Table 5
Comparison results of recognition performance using different methods.

Method	Black	Blue	Cyan	Gray	Green	Red	White	Yellow	AP
BoW + FC [1]	0.9713	0.9451	0.9787	0.8461	0.7834	0.9876	0.9414	0.9457	0.9249
Parallel CNN [9]	0.9738	0.9410	0.9645	0.8608	0.8257	0.9897	0.9666	0.9794	0.9447
SPM + kernel SVM [7]	0.9796	0.9642	0.9886	0.8686	0.8406	0.9926	0.9619	0.9787	0.9469
Proposed Method	0.9756	0.9834	0.9857	0.8884	0.8672	0.9938	0.9629	0.9759	0.9541

In summary, the reasons why the vehicle color recognition method in this paper can obtain quick and precise recognition performance are as follows.

(1) According to the requirements of the vehicle color recognition task, a lightweight convolutional neural network is specially designed, greatly simplifying the network structure and improving the recognition speed.
(2) The proposed method makes full use of the features of the intermediate layers, which can provide more supplemental information to the final global features to more efficiently describe the characteristics of the vehicle images, and thus, improving the recognition accuracy.
(3) SPM strategy is adopted to represent the feature maps of the convolutional layers compactly, which can embed the spatial information into the representation vectors to improve their descriptive and discrimination capability. Through encoding of SPM division, the feature representation capability can be enhanced while the dimension will not bring heavy burden on the computational complexity.

4. Conclusion

In this paper, a vehicle color recognition method based on lightweight convolutional neural network is proposed. Compared to the conventional methods, the proposed method in this paper enjoys two merits: the first is to design a special lightweight convolutional neural network for vehicle color recognition task, reducing the quantity of network parameters and lowering the demand for computational and storage resources during network training; the second is to apply SPM strategy to embed spatial information into convolutional feature map, combining the features from low to high layers to improve the deep feature representation capability.

In this paper, the encoding method of SPM division remains a little coarse. In the future work, new encoding method will be designed to further improve feature representation capability and recognition performance.

Acknowledgments

The work in this paper is supported by the National Natural Science Foundation of China (No. 61531006, No. 61372149, No. 61370189, No. 61471013, and No. 61602018), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions(No. CIT&TCD20150311), the Science and Technology Development Program of Beijing Education Committee (No. KM201510005004).

References

[1] P. Chen, X. Bai, W. Liu, Vehicle color recognition on urban road by feature context, Intell. Transp. Syst. IEEE Trans. 15 (5) (2014) 2340–2346.
[2] G. Qiu, K.M. Lam, Spectrally layered color indexing, Image Video Retrieval (2002) 100–107.
[3] J. Huang, S.R. Kumar, M. Mitra, et al., Image indexing using color correlograms, in: Computer Vision and Pattern Recognition, 1997. Proceedings, 1997 IEEE Computer Society Conference on, IEEE, 1997, pp. 762–768.
[4] N. Baek, S.M. Park, K.J. Kim, et al., Vehicle color classification based on the support vector machine method, in: Advanced Intelligent Computing Theories and Applications. With Aspects of Contemporary Intelligent Computing Techniques, 2007, pp. 1133–1139.
[5] E. Dule, M. Gokmen, M.S. Beratoglu, A convenient feature vector construction for vehicle color recognition, in: Proc11th WSEAS International Conference on Neural Networks, Evolutionary Computing and Fuzzy systems, 2010, pp. 250–255.
[6] W. Hu, J. Yang, L. Bai, et al., A new approach for vehicle color recognition based on specular-free image, Sixth International Conference on Machine Vision (ICMV 13), International Society for Optics and Photonics, 2013 90671Q-90671Q.
[7] C. Hu, X. Bai, L. Qi, et al., Vehicle color recognition with spatial pyramid deep learning, Intell. Transp. Syst. IEEE Trans. 16 (5) (2015) 2925–2934.
[8] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[9] Rachmadi R.F., Purnama I. Vehicle color recognition using convolutional neural network[J]. arXiv:1510.07391, 2015.
[10] Springenberg J.T., Dosovitskiy A., Brox T., et al. Striving for simplicity: Tthe all convolutional net[J]. arXiv:1412.6806, 2014.
[11] Lin M., Chen Q., Yan S. Network in network[J]. arXiv:1312.4400, 2013.
[12] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, 2, IEEE, 2006, pp. 2169–2178.
[13] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[14] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556, 2014.
[15] Y. Jia, E. Shelhamer, J. Donahue, et al., Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.