Hierarchical traffic sign recognition based on multi-feature and multi-classifier fusion

Yunxiang Ma, Linlin Huang
School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
Email: mayunxiang1988@126.com, huangll@bjtu.edu.cn

***Abstract*—In this paper, we propose a fast and robust method for traffic sign recognition, which uses a coarse-to-fine strategy. The traffic signs are divided into main category and sub-category. At the coarse classification stage, we extract histogram of oriented gradients (HOG) feature from different spectral bands of traffic sign images and classify into main category using a linear support vector machine (SVM). Then at the fine classification stage, complementary features of dense-sift, local binary pattern (LBP) and Gabor filter features are extracted, fused and then fed to a committee of SVM and random forest. The proposed method gets an accuracy of 98.76% on the German Traffic Sign Recognition Benchmark (GTSRB) dataset and takes about 50ms per image. Both recognition accuracy and speed is higher than that of the method based on multi-scale convolutional neural network.**

Keywords—traffic sign recognition; multi-feature fusion; multi-classifier fusion

I. INTRODUCTION

Traffic sign recognition (TSR) in real-word environments has important applications in advanced driver assistance system (ADAS) and self-driving technology. As an important part of intelligent transportation system (ITS), traffic sign recognition system can inform the drivers of road conditions, and hence enhance traffic safety and reduce traffic accidents.

In practice, a large number of different sign classes needs to be recognized with high accuracy. Though traffic signs have been designed to be easily readable for humans, for computers systems, classifying traffic signs still seems to be a challenging pattern recognition problem due to changes of illumination, motion blur, varying weather conditions, partial occlusions, rotation, low resolution, deterioration, and so on.

Traditionally, most traffic sign recognition systems are composed of pre-processing, feature extraction and classification. Highly distinctive image representation can boost the recognition performance. Recently many artificially designed image descriptors have been applied in traffic sign recognition, such as histogram of oriented gradients (HOG) [1] [2], scale-invariant feature transform (SIFT) [3], local binary pattern (LBP), Haar-like wavelet, shape context [4] and so on.

A few years ago, little systematic comparison of such systems exists. A publicly available traffic sign dataset with more than 50,000 images of German traffic signs in 43 classes are presented at IJCNN 2011 [2]. Many methods are proposed based on the GTSRB dataset. Ciresan et al. proposed a committee of CNNs which achieves highest accuracy of 99.46%, but suffer from huge computation cost [5]. Tang et al. employ complementary features and linear SVM and get a 98.65% recognition rate [6]. Lu et al. present a novel sparse-representation-based graph embedding algorithm that strikes a balance between local manifold structures and global discriminative information [8].

In this paper, we propose a coarse-to-fine hierarchical method based on multi-feature and multi-classifier fusion. At the coarse classification stage, we extract histogram of oriented gradients (HOG) feature from different spectral bands images and classify into super classes using linear support vector machine (SVM). At the fine classification stage, complementary features of dense-sift, local binary pattern (LBP) and Gabor filter features are extracted, fused and then fed to a committee of SVM and random forest. Experiments results show that the proposed method gets an accuracy of 98.76% on the GTSRB dataset.

This paper is organized as follows: section II introduces the overview of the proposed recognition system. Section III introduces the coarse classification stage in details. Section IV introduces the fine classification stage in details. In section V, we evaluate the proposed method on the publicly available GTSRB dataset, and compare the experimental results with other state-of-the-art methods. Conclusions and possible further improvements are given in section VI.

Fig.1. Examples for difficulties of traffic sign recognition: low resolution, rotation, motion blur, deterioration, occlusion, and bad illumination.

Fig.2. Examples of traffic signs with similar appearances

II. SYSTEM OVERVIEW

Traffic signs can be dived into different super classes according to their color, shape and interior pictograms. Traffic signs which have similar appearances are very easy to be confused. For example, some traffic signs have similar local components such as Fig.2 column 1 and 2; some traffic signs have similar global structure but different background colors, such as Fig.2 column 3; some traffic signs share the same shape and similar interior pictograms but have different pictograms' colors, such as Fig.2 column 4.

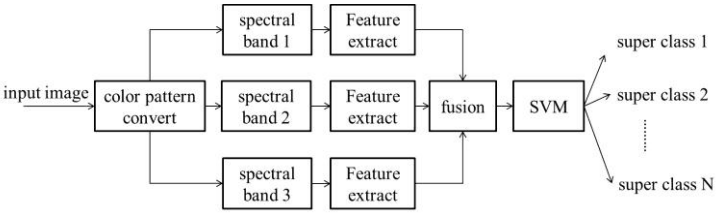


Fig.3. Flow of the coarse classification stage

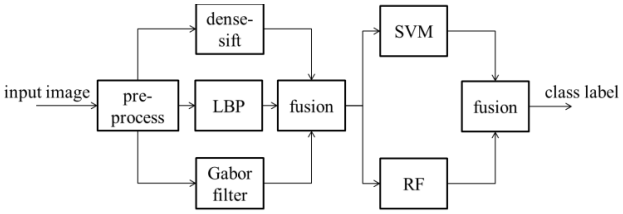


Fig.4. Flow of the fine classification stage

The proposed method consists of two stages: coarse classification and fine classification. At coarse classification stage, HOG features are extracted from different spectral bands, which are classified using a SVM, as Fig.3 shows. At the fine classification stage, multiple complementary features are extracted and fed to a committee of SVM and random forest, as Fig.4 shows.

III. CAORASE CLASSIFICATION

A. Color local image representation

Color local texture feature is introduced in [9] for face recognition and achieves good performance. The so-called color local texture features encode the discriminative features derived from different spectral channels (or bands) within a certain local region. It can make full use of both color and texture information. Considering that traffic signs have obvious color and shape, the color local texture feature could be suit for traffic sign recognition.

B. Histogram of Oriented Gradient (HOG)

HOG was proposed by Dalal and Triggs for human detection in 2005 [10], which had been widely applied in many other object detection and recognition tasks. It represents the occurrence of gradient orientations. It can capture large scale structure global shape of traffic signs.

A Sobel filter is used to find the horizontal and vertical derivatives and the magnitude and orientation for each pixel, which are computed as equation (1) ~ (4).

We find that the application of HOG to traffic sign recognition very suitable, given that traffic signs are composed of strong geometric shapes and high-contrast edges that encompass a range of orientations.

$$G_x(x,y) = H(x+1,y) - H(x-1,y) \quad (1)$$

$$G_y(x,y) = H(x,y+1) - H(x,y-1) \quad (2)$$

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \quad (3)$$

$$\alpha(x,y) = \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \quad (4)$$

In this paper, we divide the images into 6×6 blocks. Each block has 2×2 cells, and each cell has 12×12 pixels. The number of orientation bins is 8.

C. Coarse classification

In this paper, we extract HOG features from different spectral bands. Several color space are used, such as RGB, HSV, XYZ, Lab, and some artificial color patterns e.g. color angle pattern.

For coarse classification, we use support vector machine (SVM) as the classifier. The theory about SVM is briefly introduced below. The extracted color HOG features are fed to a one-vs-all SVM.

IV. FINE CLASSIFICATION

A. Preprocessing

Traffic signs in the same super class have same background color, same shape and similar icons, so gray-scale images are enough. Firstly, the input color images should be converted into gray-scale images. In order to reduce the illumination's influence, then, intensity normalization with the mean intensity 180 and variance 30 is implemented on the gray-scale images. At last, the images are resized to 48×48 pixels using bilinear interpolation algorithm. It's worth noting that in the danger signs subset, general danger sign and traffic light sign's icons are similar but have different colors. So color images are used on danger signs when extracting features.

B. Multiple complementary local features

Recently, local textures feature have gained reputation as powerful descriptors because they are believed to be more robust to variations of rotation, occlusion, etc.

1) Dense SIFT

Scale-invariant feature transform (SIFT) is proposed by Lowe [11]. The SIFT features are invariant to image scale and rotation.

Instead of extracting descriptors around interest points only, local feature descriptors are extracted at regular image grid points that allow for a dense description of the traffic sign images

In general, dense-sift is very similar to HOG. In this paper, the images are divided into 12×12 cells with 4×4 pixels per cell.

2) Local binary pattern (LBP)

Local binary pattern is first proposed by Ojala et al. for texture classification [12]. The original LBP operator is calculated in a local 3×3 neighborhood, comparing the center pixel value with its eight neighborhood pixels: if the neighbor pixel value is bigger than the center pixel value, then, the neighbor pixel is labeled as 1, otherwise, labeled as 0, as equation (5) and (6). Finally, the center pixel gets a LBP code, as the Fig.6 shows. We divide the input traffic sign images into non-overlapping blocks. The histograms of each block are calculated and concatenated to form the LBP feature vector.

$$LBP_{P,R}(x_c, x_y) = \sum_{P=0}^{P-1} s(i_p - i_c)2^P \quad (5)$$

$$s(x) = \begin{cases} 1, & if \quad x \geq 0 \\ 0, & if \quad x < 0 \end{cases} \quad (6)$$

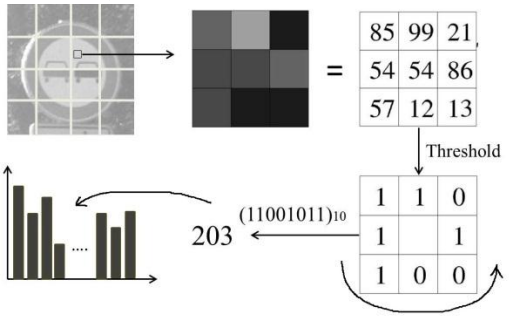


Fig.5. Flow of the original LBP operator

In order to reduce the dimension of LBP descriptor, we employ the so-called uniform patterns to reduce the

bins of histogram of each block from 256 to 59. In this paper, we divide the images into 6×6 non-overlapping blocks, namely 8×8 pixels per block.

3) Multi-scale Gabor filter feature

Gabor filters with low orientation sensitivity and broad frequency band favor recognition accuracy.

$$g(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\frac{1}{2}\left(\frac{\tilde{x}^2}{\sigma_x^{\,2}}+\frac{\tilde{y}^2}{\sigma_y^{\,2}}\right)\right]\exp(i2\pi f\tilde{x}) \quad (7)$$

$$\tilde{x}=x\cos\theta+y\sin\theta \quad (8)$$

$$\tilde{y}=-x\sin\theta+y\cos\theta \quad (9)$$

We use the magnitude as the final Gabor filter feature. In this paper, we divide the images into 8×8 overlapping blocks. Each block contains 12×12 pixels. The Gabor filters are designed with 2 scales and 8 orientations.

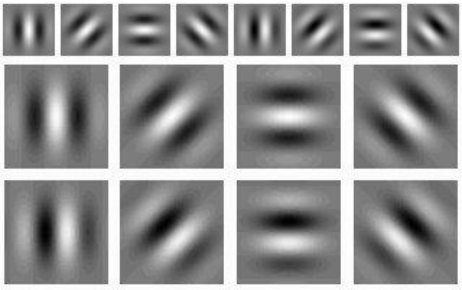


Fig.6. Real and imaginary parts of Gabor filters with 4 orientations and 2 scales (left for real parts and right for imaginary parts)

C. Multi-feature fusion

Generally speaking, techniques for fusing multiple evidences (multiple classification results or multiple features) can be divided into two classes, i.e., fusion at the "feature level" and fusion at the "decision level". In this paper, we use the feature-level fusion strategy, since "feature-level fusion" can generally achieve a better classification result compared with "decision-level fusion".

The feature could be simply concatenated into a longer global feature vector. Then, several complementary low-dimensional features are combined at the level of the features by concatenating low-dimensional features in the column order. The three descriptors are complementary for traffic sign recognition. We compute the compound features by equation (10).

$$f_I=\left[(f_{SIFT})^T,(f_{LBP})^T,(f_{Gabor})^T\right]^T \quad (10)$$

D. Multi-classifier fusion

Generally, SVM classification is fast, highly accurate, and less prone to overfitting compared to many other classification methods. Random forest performs comparatively to SVM and outperforms naïve Bayes, KNN, and C4.5 classifier. The main idea of random forest is grow an ensemble of decision trees and allowing them to vote for the most popular class.

To test the performance of the random forest, we vary some parameters e.g. the decision trees number, the number of features chosen at random.

In this paper, we choose 500 trees to form a random forest. The training sets are randomly chosen from the whole GTSRB dataset for each tree.

The SVM outputs a score for each class, denoting the confidence value of prediction. Denote the SVM's output score as m_i, $i=1,2,3...k$. Denote the random forest's output votes as n_i, $i=1,2,3...k$.

The outputs are normalized as equation (11) and (12).

Add the SVM's output and random forest output together, as equation (13), the largest label is the final class of input traffic sign image.

$$\tilde{m}_i=\frac{m_i}{\sum_{i=1}^{k}m_i},\quad i=1,2,3\cdots k \quad (11)$$

$$\tilde{n}_i=\frac{n_i}{n},\quad i=1,2,3\cdots k \quad (12)$$

$$class=\max(\tilde{m}_i+\tilde{n}_i),\quad i=1,2,3\cdots k \quad (13)$$

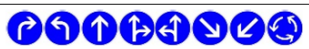
V. Experimental Results

In order to evaluate the effectiveness of our proposed method, we use the publicly available German traffic sign recognition benchmark (GTSRB) dataset. The experiments results include two parts, namely, coarse classification and fine classification.

A. Dataset

The GTSRB dataset was produced for the IJCNN2011 competition and released for public use. It contains 51,839 traffic signs images in 43 classes including 39,209 training images and 12,630 test images. The images only contain one traffic sign each and contain a border of 10% around the actual traffic sign (at least 5 pixels) to allow for edge-based approaches. The image sizes vary between 15×15 and 250×250. Additionally the images are not necessarily squared. Table I shows the 43 classes of standard traffic signs in GTSRB dataset, which have been divided into six subsets: *speed limit signs*, *other prohibitory signs*, *derestriction signs*, *mandatory signs*, *danger signs*, and *unique signs*.

TABLE I
SUPER CLASS OF GERMAN TRAFFIC SIGNS

Subset	standard traffic signs
(a) Speed limit signs	
(b) other prohibitory signs	
(c) Derestriction signs	
(d) Mandatory signs	
(e) Danger signs	
(f) Unique signs	

B. Results

We performed the proposed method based on the platform of a desktop computer with a Core E8400 (3.0Hz) and 2GB DDR3 and VS2010.

1) coarse classification results

We get a best accuracy of 99.79% on the whole GTSRB dataset. As TABLE II shows, the linear SVM gets the best performance in the coarse classification stage. The linear SVM not only achieves high accuracy but also costs least compared with other kernel function.

From TABLE III, we can notice that the RGB color space performs best, achieving a 99.79% recognition rate. So we should choose RGB color space when extracting the color local feature.

1) fine classification results

As we can see from TABLE IV, the fusion of the above three features outperform every single feature,

proving that different features are complementary. In addition, a one-versus-all SVM achieves a 98.51% recognition rate, and random forest achieves accuracy of 97.94%. What's exciting is that the committee of SVM and random forest boost the accuracy to 98.76%, which is close to human performance.

C. Discussion

The combination of three features outperforms a signal feature, and the committee of SVM and RF outperform a signal classifier. The two strategies both boost the recognition performance.

Most of the misclassified images are caused by low resolution and bad illumination. Some images are even difficult for human. Additionally we compare our method with other methods based on the GTSRB dataset, as TABLE VI shows. The committee of CNNs achieves the best recognition rate of 99.46%, much better than human performance. But the committee of CNNs cost too much both training and test, not meeting the requirement of real-time application.

TABLE II
RECOGNITION RESULTS WITH DIFFERENT KERNEL FUNCTIONS

Kernel function	Training time (s)	Test time (s)	Accuracy (%)
Liner	0.04	0.02	99.79
RBF	3.84	0.22	99.52
Polynomial	7.62	0.75	98.96
Sigmoid	4.93	0.30	95.63

TABLE III
RECOGNITION ACCURACY UNDER DIFFERENT COLOR SPACE

Color space	Accuracy (%)
RGB	99.79
XYZ	99.63
YCrBr	99.54
Lab	99.52
HSV	99.48
gray-scale	98.95
color angle pattern	97.86

TABLE IV
COMPARISON BETWEEN SIGNAL FEATURE AND COMPOUND FEATURE USING LINER SVM

feature	Accuracy (%)
Dense-sift	97.28
LBP	96.94
Gabor filter	97.09
Multi-feature fusion	98.61

VI. CONCLUSIONS

In this paper, we propose a fast and robust hierarchical method with high accuracy based on multi-feature and multi-classifier fusion. Experimental results show that our proposed method achieves a 98.76% recognition rate on the GTSRB dataset, which is very close to the average human performance. The fusion of complementary features and the committee of SVM and RF boost the recognition performance.

TABLE V
COMPARISON BETWEEN SIGNAL CLASSIFIER AND COMPOUND CLASSIFIER ON COMPOUND FEATURE

classifier	Accuracy (%)
SVM	98.51
RF	97.94
SVM+RF	98.76

TABLE VI
COMPARISON BETWEEN THE PROPOSED METHOD AND OTHER METHODS

Methods	Accuracy (%)
Committee of CNNs[5]	99.46
Human performance(average)[2]	98.84
The proposed method	**98.76**
Complementary features[6]	98.65
Multi-scale CNN[7]	98.31
SRGE[8]	98.19
Random forests[13]	96.14
LDA on HOG2[2]	95.68

ACKNOWLEDGMENT

This work was sponsored in part by the National Natural Science Foundation of China with Grant No.63271306. The authors would like to thank Suisui Tang for his help.

REFERENCES

[1] Greenhalgh, Jack, and Majid Mirmehdi. "Real-time detection and recognition of road traffic signs." Intelligent Transportation Systems, IEEE Transactions on 13.4 (2012): 1498-1506.

[2] Stallkamp J, Schlipsing M, Salmen J, et al. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition [J]. Neural networks, 2012, 32: 323-332.

[3] Kus, M. C., M. Gokmen, and S. Etaner-Uyar. "Traffic sign recognition using Scale Invariant Feature Transform and color classification." Computer and Information Sciences, 2008. ISCIS 08. 23rd International Symposium on. IEEE, 2008.

[4] Jin Qin, Xinfeng Zhang. Hierarchical traffic sign recognition system based on improved shape context [J]. Computer Engineering and Design, 2014, 35(1): 183-187.

[5] Ciresan D, Meier U, Masci J, et al. A committee of neural networks for traffic sign classification[C]//Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011: 1918-1921.

[6] Tang S, Huang L L. Traffic Sign Recognition Using Complementary Features[C]//Pattern Recognition (ACPR), 2013 2nd IAPR Asian Conference on. IEEE, 2013: 210-214.

[7] Sermanet P, LeCun Y. Traffic sign recognition with multi-scale convolutional networks[C]//Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011: 2809-2813.

[8] Lu K, Ding Z, Ge S. Sparse-representation-based graph embedding for traffic sign recognition [J]. Intelligent Transportation Systems, IEEE Transactions on, 2012, 13(4): 1515-1524.

[9] Choi, Jae Young, Yong Man Ro, and Konstantinos N. Plataniotis. "Color local texture features for color face recognition." Image Processing, IEEE Transactions on 21.3 (2012): 1366-1380.

[10]Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[11] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International journal of computer vision, 2004, 60(2): 91-110.

[12] Ahonen, Timo, Abdenour Hadid, and Matti Pietikainen. "Face description with local binary patterns: Application to face recognition," Pattern Analysis and Machine Intelligence, IEEE Transactions on 28.12 (2006): 2037-2041.

[13]Zaklouta F, Stanciulescu B. Real-time traffic-sign recognition using tree classifiers [J]. Intelligent Transportation Systems, IEEE Transactions on, 2012, 13(4): 1507-1514.