Extreme Learning Machine with Spatial and Time-Frequency Feature Fusion for High Performance Traffic Sign Recognition

Xiaoxiao Zhou[1], Zhi-Xin Yang[1,⋆], and Peng-bo Zhang[2]

[1]Department of Electromechanical Engineering, Faculty of Science and Technology, University of Macau, Macau
[2]Department of Industrial Engineering and Logistics Management, school of Engineering, Hong Kong University of Science and Technology, Hong Kong

Abstract. Extremely high accuracy and efficiency of Traffic Sign Recognition (TSR) are crucial for driving autonomous vehicles(AVs). A new extreme learning machine based TSR framework with spatial and time-frequency feature fusion technology, named as STF-ELM, is proposed in this paper. Histogram equalization is employed to enhance the traffic sign image preprocessing, which contributes to balancing distribution of intensity and sharpness. In addition to the existing histogram of oriented gradients(HOG) for extracting spatial features, the time-frequency domain's feature, Gabor representation, is also used. The compressively fused feature vector is mapped to the predicated traffic signs via ELM network. The experimental results showed that the proposed STF-ELM achieved almost perfect performance with zero-error for most of categories, and outperformed the state-of-the-art TSR methods on the German Traffic Sign Recognition Benchmark(GTSRB).

Keywords: traffic sign recognition, extreme learning machine, histogram equalization, feature fusion, histogram of oriented gradients, gabor

1 Introduction

Traffic sign recognition(TSR) is a technology which helps vehicles recognize upcoming traffic signs on the road while driving. Specifically, it plays an essential role in autonomous vehicles(AVs), which are designed to drive automatically without drivers. AVs is a very promising innovative area which upends automotive industry, for the reason that human beings are inevitable to miss traffic signs while driving, and this would result traffic accidents public safety threats. AVs are equipped with advanced camera system to scan roads and detect traffic signs as well as surroundings, then recognize them by TSR technology while driving, taking place of human manipulation. As a result, AVs make instant decisions by itself, and then could overcome the potential hazards of drivers' unavoidable

⋆ Corresponding author. Email address: zxyang@umac.mo

negligence. Therefore, as absolute security is required for driverless transportation, the extremely high recognition accuracy and extremely high efficiency of TSR are two critical factors for AVs. Despite only one small fault with a traffic sign, such as mistaking 'left turn' as 'right turn', it will certainly cause great traffic disturbance and most likely to trigger accidents.

However, TSR with perfect recognition accuracy and efficiency has been a tricky and challenging problem for decades. There are three challenges mainly. The first lies in poor-quality images collected in the dynamic driving process, caused by disturbance, like bad weather, occlusions, motion blur, viewpoint variations, too bright or too dark illumination [1]; it is hard to extract robust features from these images. Compared with poor images, clear and high-contrast images have recognizable shapes of signs, then shallow but representative features could be easily abstracted from them and recognized by TSR. Thus, how to overcome objective inevitable disturbance and improve these images is of great concern.

The second difficulty of TSR is how to get representative and robust features from images. Just as mentioned above, good performance of classification heavily depends on good images quality and discriminated features. In many related work, handcrafted descriptors are well adopted, including histogram of oriented gradients(HOG) [2], Gabor features [3], speeded up robust features(SURF) [4], scale-invariant feature transform(SIFT) [5] [6], and texture features [7] , etc. They are manually designed for specific datasets firstly, and used for extracting shallow features. SIFT and SURF are two different local detectors and descriptors with good accuracy. But SIFT is slow and not quite good at illumination variations. Althoug SURF is faster, it is weak on rotation and brightness changes relatively [8]. HOG and Gabor are both sensitive to edges and invariant to disturbance. HOG takes advantages of sharp intensity changes of edges, calculating the gradients in different directions to detect the shapes [9], but it is sensitive to noises because of gradients. Gabor uses window-Fourier transform to compute intensity distribution and reduce noises by its linear filter [10]. Each feature descriptor has its weakness for TSR. As discussed, how to design and extract robust and efficient features for TSR is the second challenge.

In the progress of TSR, classifiers play a crucial role in recognition performance. Shallow classifiers have been taken full advantage of its simple structure and high training efficiency for years, including support vector machines(SVM) [11], adaboost, K-d trees, random forests [12], and some ensemble methods [13] [14] etc. For simple datasets, shallow classifiers is quite efficient, but for complicated images, they performs relatively poor; especially in our case, images are collected while moving. Recently, various state-of-the-art deep neural networks(DNN) have been put forward for multi-class classification [15] [16]. DNN is a multi-layer feedforward network with many hidden nodes, trained by back propagation(BP) algorithm and its network weights are adjusted by calculating gradients iteratively. Although iterations benefit high training accuracy, it is likely to get a local minimum or over-fitted with high computational and time costs [17]. Therefore, designing a classifier achieves a good balance between recognition accuracy and efficiency is the third challenging problem in TSR.

In respect of the three issues above, we proposed a fast and precise model known as based Extreme Learning Machine Spatial and Time-Frequency Feature Fusion(STF-ELM), utilizing fusion of spatial HOG and time-frequent Gabor features as hybrid feature, which is robust to different variations and sensitive to edges, and employing extreme learning machine(ELM), which performs excellent in prediction accuracy at high learning speed [18]. Our method is implemented on the German Traffic Sign Recognition Benchmark(GTSRB), which is set up by Stallkamp *et al.* [19]. The details about GTSRB will be discussed in following contents.

This paper is organized as follows. Section 2 discusses previous work on TSR. Section 3 presents the overall framework and details of our proposed STF-ELM model. Section 4 discusses experiments comparisons and results. Section 5 verifies STF-ELM and concludes.

2 Related work

Great efforts have been devoted to traffic sign recognition(TSR) within past two decades. The basic procedures of TSR have three stages: Image preprocessing, feature extraction, and classification.

2.1 Image preprocessing

Distinguished recognition performance is inextricably bound up the input image quality, which mainly depends on shape, or color information of signs. In consideration of specific practical application for AVs' real-time TSR systems, image preprocessing methods are essential for further feature extraction [20] [7]. Rich color information are commonly used as basis of models, but it is severely affected by illumination, bad weather conditions, biased points of view, etc [21]. So before that, color space conversion are commonly used to overcome these issues, such as Lab [22], YUV [23], HSV, etc. Although color is informative, it brings much noise from background as well. Many models mainly focus on shape information, and only high contrast images have clear shapes. Contrast enhancement methods are required to modify distribution of intensity uniformly [24] to get distinctive borders of traffic signs. For example, Cireşan *et al.* [25] proposed a committee of neural networks, in which they utilized Contrast-limited Adaptive Histogram Equalization(CLAHE) to divide the image into several non-overlapping regions of almost equal size. Jin *et al.* [15] used three types of histogram equalization(HE) methods, including global histogram equalization(GHE), adjusted image intensity values and CLAHE for normalizing images and achieve better contrast. However, it is undeniable that just a fringe of models employed HE methods, and most of models supposed usage of anti-interference descriptors would be enough for recognition. In fact, it is hard for those descriptors to get better representative features from blurred images with uneven attributed of gray levels.

2.2 Feature extraction

A robust and discriminate feature descriptor is worthwhile to design, for its competitive ability of abstracting informative and representative features. As addressed previously, handcrafted descriptors are frequently used. HOG descriptor was firstly put forward by Dalal Navneet and Triggs Bill for human detection in 2005 [9], which is proved in the following years to be very effective for invariance to disturbance, and then widely used in various domains. Wang *et al.* [11] proposed a two-layer model; both of layers are designed with HOG descriptor and support vector machine(SVM) for between-category and within-category classification. This method achieved overall accuracy of 99.52%. Huang *et al.* [26] put forward an optimized HOG named HOG variant(HOGv), including signed and unsigned gradients. Also, there are newly designed descriptors. Yuan *et al.* [21] came up a novel Color Global and Local Oriented Edge Magnitude Pattern(Color Global LOEMP), which was robust to brightness, rotation and scale, achieved an accuracy of 97.2581%. Anjan *et al.* [7] designed a feature combined with higher order spectra(HOG) and texture features, and achieved overall accuracy of 98.89%.

2.3 Classification

For TSR, the state-of-the-art CNN framework is widely employed by many models for classification. Zhu *et al.* [16] proposed a comprehensive system with detection and classification based on a completely new dataset collected by themselves. The team trained two CNNs, one for detecting and the other for classification with accuracy of 88%. Zeng *et al.* [22] came up with a cascaded neural network with two CNNs, of which the first is for reducing different images and the second is trained to recognize low-contrast images. This hierarchy network achieved an accuracy of 99.45%, enhancing precision compared with traditional CNN, but the computational cost is extremely expensive with high demand of computers configuration. In consideration of practical application, DNN can hardly satisfy demands of real-time driving system for detecting and recognizing traffic signs while driving.

In recent years, a single hidden layer feed-forward neural network(SLFN) named extreme learning machine(ELM) is proposed [18] which randomly chooses hidden nodes and analytically determines the output weights of SLFN instead of iterated constantly by BP. Many algorithms based on ELM are proposed [27] [28]. Wong *et al.* [29] proposed a modified multi-layer extreme learning machine with kernel(ML-KELM). Wong *et al.* [30] integrates feature extraction, parameter optimization algorithm, and multiple sparse Bayesian extreme learning machines (SBELM) as an intelligent diagnostic framework. Huang *et al.* [26] combine signed and unsigned HOG as feature and employ the ELM as classifier tested on three datasets including GTSRB, and accuracy of 98.38% improves better generalization and outperforms most of state-of-the-art algorithms for accuracy and efficiency.

3 Proposed TSR Frameworks

This proposed TSR framework consists of three main submodules: Traffic Sign image preprocessing, feature extraction and fusion, and classfication. The architecture of STF-ELM is depicted in Fig 1.

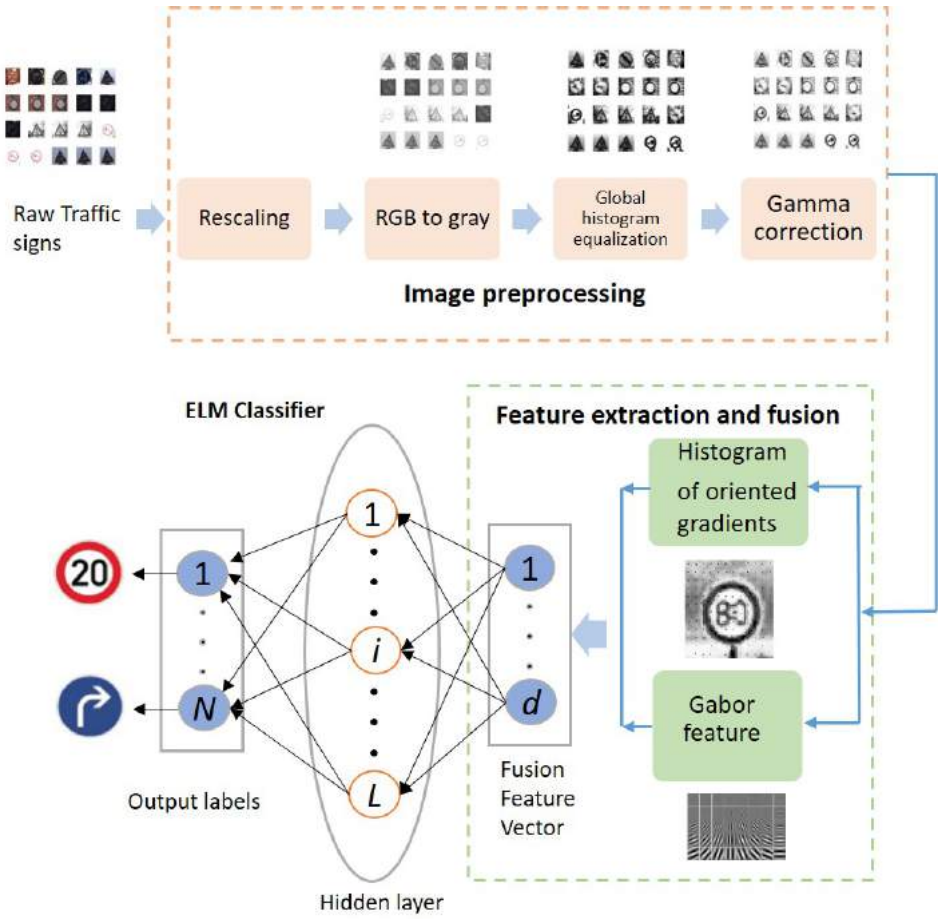


Fig. 1: The architecture of STF-ELM

3.1 Image Preprocessing

For the various size, the first step is rescaling images to a proper size using bilinear interpolation and converting them to gray. Bilinear interpolation considers the closest 2×2 neighborhood of known pixel values to calculate the interpolated

value of the middle unknown pixel, for this reason, images could still keep large details. But still, most of them are blurred or predominantly dark or bright, which could hardly be recognized even by eyes. Some samples of TR images from GTSRB are shown in Fig 2.

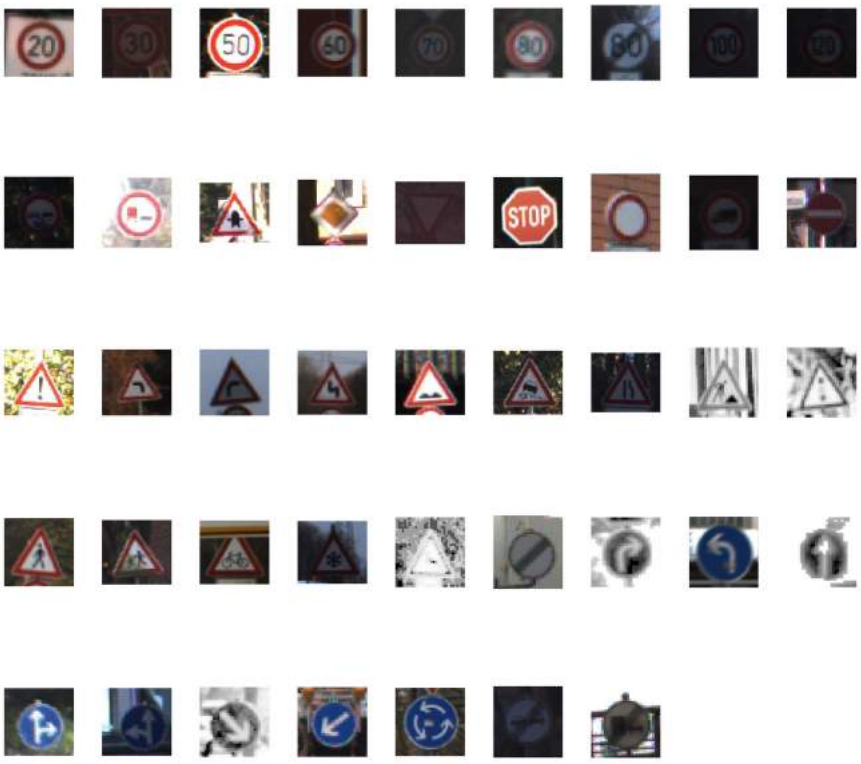


Fig. 2: 43 classes of typical traffic signs in the GTSRB database

Hence, in this study, Global histogram equalization(GHE) is employed to enhance contrast and make clear. From the observation of GTSRB images in Fig 2, we find that usually the intensity varies in a similar gray level range over a whole image, so the global histogram equalization(GHE) is suitable for GTSRB with overall enhancement.

GHE is based on the cumulative density function(CDF) to stretch intensity to distribute uniformly, which can remain the density distribution of intensity unchanged after stretch. Given an image X of width m and height n. $X_{m\times n}$ has $m \times n$ pixels in the range of K_1 to K_2. Suppose P_i is the occurrence of each pixel

i, and the CDF of each pixel i is defined as

$$F(i) = P(x <= i) = \sum_{x=K_1}^{i} P_i, K_1 =< i =< K_2 \tag{1}$$

The minimum of F(i) is F_{min}, and the maximum is $m \times n$. The CDF is to be mapped into C_1 to C_2. So the general GHE formula is:

$$h(i) = round(\frac{F(i) - F_{min}}{m \times n - F_{min}} \times (C_1 - C_2)) \tag{2}$$

For our TSR model, the traffic signs images are scaled to 48×48 in gray and mapped into 0 to 255 by HE. The above equation would be:

$$h(i) = round(\frac{F(i) - F_{min}}{48 \times 48 - F_{min}} \times (255 - 0)) \tag{3}$$

As an example, the result of GHE is illustrated in Fig 3 . We can find that the original image is too dark to identify what it is, but GHE clarified it.

3.2 Feature extraction and fusion

Traffic sign recognition is achieved by distinguishing features of objects, color, frequency, spatial information, etc. In this study, histogram of oriented gradients(HOG), and Gabor features are employed respectively for space and time-frequency features, and HOG and Gabor features are fused as a feature for TSR.

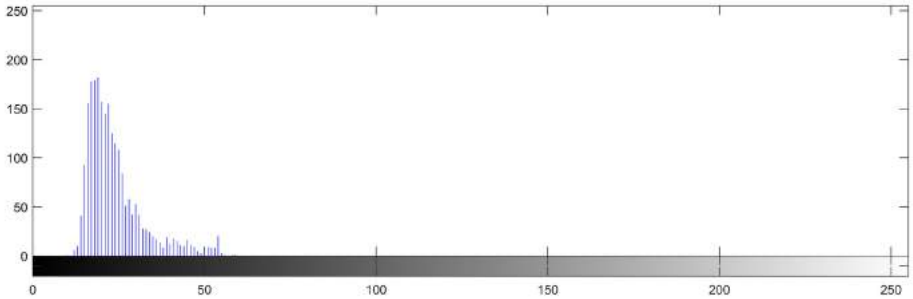
HOG features HOG uses oriented gradients distribution as features. Locally, borders and corners usually have sharp intensity changes, so large-value gradients gather around them and visualized as Fig 4 and Fig 5.

This benefits TSR to detect edges of traffic signs. Furthermore, local normalization within each block is computed in HOG, which makes HOG be invariant to lighting variations. From Fig 4 - 5, it is obvious that HOG have clearly drawn the outline of the speed limit sign. But just for its calculating gradients, HOG is sensitive to noises.

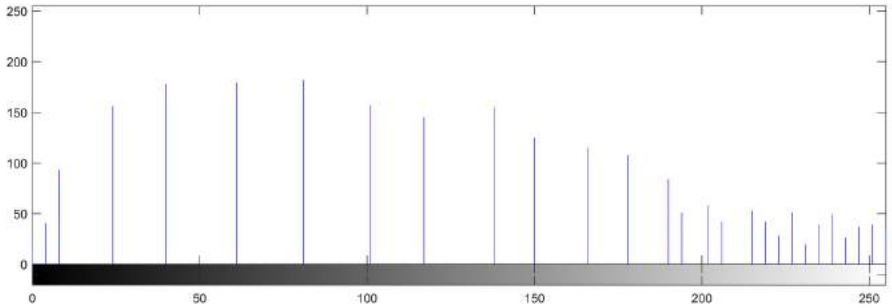
Gabor features Gabor is a linear filter based on time-frequency domain using window Fourier transform, which is employed in this study to extract edge information [31]. The 'window' function helps Gabor learn local details at any partial time, and the frequency of the whole image indicates the extent of gray variation integrally. Besides, Gabor is anti-noise for its filters. Therefore, Gabor features can provide overall frequency map and local time-dependent properties of the images.

(a) The original TR

(b) The GHE processed TR



(c) Gray level histogram of the original TS



(d) Gray level histogram of the GHE processed TS

Fig. 3: Comparisons of Traffic Sign (TR) pre-processing with and without GHE

Fig. 4: A preprocessed speed limit sign of GTSRB

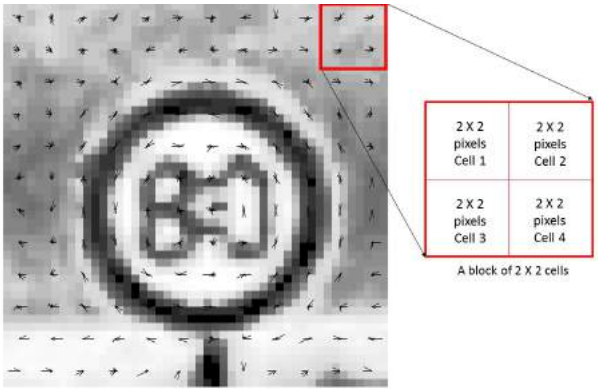


Fig. 5: A visualized HOG feature on the processed speed limit sign of GTSRB

Feature fusion We fuse HOG feature and Gabor feature for information of space domain and frequency domain. They are both sensitive to borders of objects and invariant to illumination. Their combination contributes to offering comprehensive features for classifiers and enhancing robust to noises of images.

Feature fusion is integration of several relevant data representations, indicating more accurate and informative descriptions of images than individual [32]. If we just concatenate HOG and Gabor features together, the dimension of the combined features should be too high that may cause data redundancy and increase computational costs.

Principal components analysis(PCA) is a classical algorithm for revealing the internal structure [33]. It is remarkable for reducing dimensions and extracting the most representative components which could re-express the complex images [34]. Therefore, STF-ELM employs PCA to fuse HOG and Gabor features.

The process of STF-ELM feature fusion is described as in **Algorithm 1**:

Spatially extracted HOG feature is designed to detect edges of objects and draw the shape. Gabor feature based on time-frequency domain is employed to achieve frequency distribution of images and time based local details. As a result, feature fusion attributes to enriching expression of images.

3.3 Extreme learning machine

Extreme learning machine(ELM) is a simple but efficient and highly accurate algorithm, which is first proposed by Huang *et al.* [18] based on generalized single-hidden layer feedforward neural network(SLFN), which is expert in overcoming the issues of back propagation(BP) networks, time-consuming, overfitting, local minimum and dependence on adjusting parameters.

As aforementioned, the output $\mathbf{Y}_{k\times N}$ of feature fusion section is the feature matrix of the dataset, as well as the input of classifier ELM. The training process of ELM is summarized as in **Algorithm 2** [27]:

Algorithm 1 Feature extraction and fusion algorithm in STF-ELM

1. Suppose that STF-ELM is implemented on a dataset of N images. Given an image X of the dataset, with a size of $i \times j$ pixels, as the input of HOG feature descriptor, and it is divided into $c_1 \times c_2$ none-overlapping cells. Suppose we calculate m oriented gradients in a cell, then we group the cells into overlapped blocks of $b_1 \times b_2$ cells per block with the overlap of $p_1 \times p_2$ cells. So the stride would be $((b_1 - p_1) \times c_1) \times ((b_2 - p_2) \times c_2)$ pixels.
2. Then HOG descriptor is operated on the image X to calculate gradients in n directions in each cell as a cell feature with the dimension of n, then features of $b_1 \times b_2$ cells are combined as a block feature with $n \times b_1 \times b_2$. After overlapped calculation, the features of each block are combined as a feature vector $\mathbf{Vh}_(d1 \times 1)$, the dimension d1 of which could be calculated as:

$$d1 = n \times b_1 \times b_2 \times (\frac{i - b_1 \times c_1}{(b_1 - n) \times c_1} + 1) \times (\frac{i - b_2 \times c_2}{(b_2 - n) \times c_2} + 1) \tag{4}$$

3. After HOG extracted, the Gabor descriptor is implemented on the image X to extract represents of it based on time-spatial domain, which is a feature vector as $\mathbf{Vg}_(d2 \times 1)$, with the d2 dimensions.
4. Then, the HOG and Gabor features are concatenated as $\mathbf{V}_d$ as the feature of image X, with a extremely high dimension of $d = (d1 + d2) \times 1$.
5. After all images of dataset processed, the concatenated feature of each image is combined in a big matrix $\mathbf{V}$, as the feature expression of the dataset. Thus, the dimension of $\mathbf{V}$ is $d \times N$, which is so high that would introduce noises and increase computational costs. Thus, PCA algorithm is implemented on $\mathbf{V}$.
6. Feature of each image is normalized by subtracting off the mean of $\mathbf{V}_{i,1}$ of $\mathbf{V}$.
7. Then, calculate the eigenvectors and eigenvalues of SVD of $\mathbf{V}$.
8. Rearrange the eigenvectors and eigenvalues in order of decreasing eigenvalue, and compute the cumulative eigenvalues for each eigenvector.
9. Select a subset as matrix P of eigenvectors according to the value K of energy you select, where $0 \leqslant K \leqslant 1$.
10. Then the d dimensions of a feature is reduced to k dimensions, and $\mathbf{Y}_{k \times N} = P \times X$ is the feature matrix of the dataset after PCA, as the input of classifier.

Algorithm 2: Training of STF-ELM's classifier

1. With the label vector $\mathbf{T}_{1\times N}$ of the dataset, we suppose $(\mathbf{y}_i, \mathbf{t}_i), i = 1, \ldots, N$ is the input of ELM.
2. For the input layer, the input weights and bias $(\mathbf{w}_i, b_i)$ are randomly initialized, $i = 1, \ldots, N$
3. For the hidden layer, we assume the activation function is chosen as $g(x)$, and the output $\mathbf{H}$ of L hidden nodes can be denoted as

$$\mathbf{H} = \mathbf{h}(\mathbf{y}) = g(\mathbf{w_i y_i} + b_i), i = 1, \ldots, N \tag{5}$$

4. For the output layer, the output $\mathbf{L}$ of ELM, can be written as:

$$\mathbf{H}\beta = \mathbf{L} \tag{6}$$

   where β is the output weights.
5. Calculate β and achieve the training model.

The training process aims to minimize training error and the norm of output weights:

$$Minimize : \|\beta\| \quad \text{and} \quad \theta\|\mathbf{H}\beta - \mathbf{T}\|^2 \tag{7}$$

where θ is the learning rate. Minimizing norm of β denotes the largest distance of the separating margin of two different classes of the ELM.

Many efficient methods could be used to calculate the output weight β [35] [26]. The solution for this problem can be written as:

$$\begin{aligned} \beta &= \mathbf{H}^T(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{T}, \quad \text{when} \quad N \leq L \\ \beta &= (\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}^T\mathbf{T}, \quad \text{when} \quad N > L \end{aligned} \tag{8}$$

where $T = \mathbf{t}_1 \ldots \mathbf{t}_N$.

4 Experiments and evaluations

4.1 Dataset

This study is implemented on the GTSRB dataset [19], containing 43 classes with the sum of more than 50,000 images. All these traffic signs are collected from a ten-hour video recorded by a car while driving on the roads. Image size varies from 15×15 to 250×250 pixels, existing poor-quality images of poor lighted, motion blurred, low resolution, partial occlusions, and rotations etc. The examples of GTSRB are shown in Fig 2.

Despite challenges by low quality of signs, they simulate real environment AVs will face, and that is what makes our study meaningful and significant. On account of unbalanced number of different classes, this dataset is split into

six super categories as depicted in Fig 6, including speed limits signs, other prohibitory signs, derestriction signs, mandatory signs, unique signs, and danger signs.

Fig. 6: The super categories of GTSRB

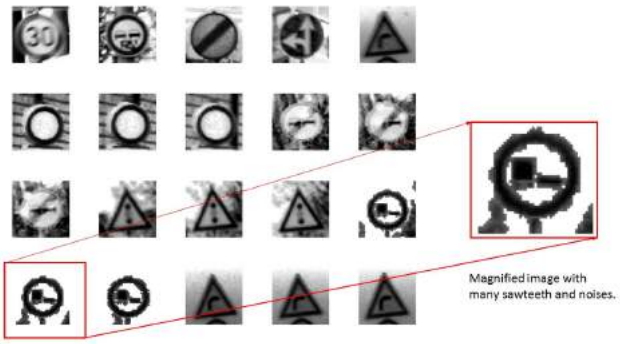
4.2 Image preprocessing setup

In this dataset, images are transformed from different size to the same size 48×48 using bilinear interpolation, and convert RGB to gray. Then, GHE should be operated on the gray images temperately. The processing procedures could be summarized as follows:

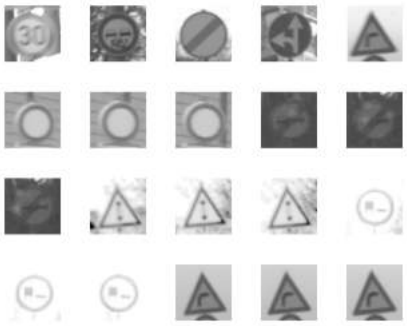
1. Rescale images to unified size 48×48 using bilinear interpolation.
2. Convert RGB to gray images.
3. Operating GHE on gray images.
4. Using gamma correction to rectify brightness.

The preprocessed images can be referred to Fig 7. It is obvious that GHE has improved images quality and make the over-dark and over-bright images clearer. Compared with (b) and (d) in Fig 7, it may be confused that the usage of GHE simply in (b), seems to make images clearer than STF-ELM preprocessed in (d) on the compressed image. However, actually, the shapes of signs in (b) are introduced with a lot of sawteeth as well as much noise. But images in (d), prove that STF-ELM not only makes images clearer, but reduce noises and keep more details than other processing methods as well.

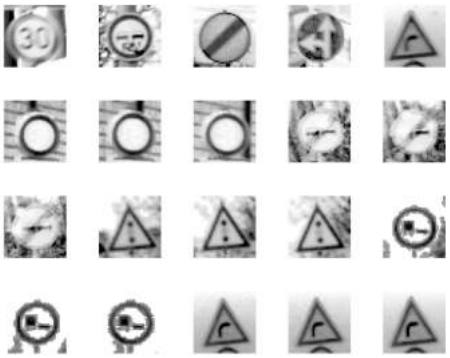
(a) Part original images of GTSRB



(b) Preprocessed images of GTSRB after GHE



(c) Preprocessed images of GTSRB using STF-ELM without GHE



(d) Preprocessed images of GTSRB using STF-ELM with GHE

Fig. 7: Comparisons of using GHE and without GHE on various images of GTSRB

4.3 Setup of feature extraction and fusion

As depicted in **Algorithm 1**, to extract HOG, the 48×48 images are divided into non-overlapping 12×12 cells with the size 4×4 pixels. In each cell, signed gradients are calculated in 14 orientation bins over 360 degrees. Then we group the cells into overlapping blocks of 2×2 cells each with 4 pixels block stride. Overlapped blocks suggest our cells are calculated four times within one image, which offer more detailed information. As a consequence, the dimension of one image equals to $14 \times (\frac{48-4}{4}) \times (\frac{48-4}{4}) \times 4 = 6776$.For Gabor features, we set five scales of Gabor filters at eight orientations, so we get 40 filters, which make high dimension of 4400 as well.

If we just concatenate two features, the concatenated feature vectors will reach as high as 11176 dimensions with much redundant information and high computational costs. With the fusion of this model, the dimensions of the feature vectors are reduced from 11176 to 895, which has been largely reduced.

4.4 Evaluation of Global Histogram Equalization

Image preprocessing plays a vital role in recognition system. While driving, this dynamic images collection is easy to achieve poor images. Therefore, in this study, GHE is used to clarify these images by enhancing contrast and balancing distribution of gray levels. In order to testify the effect of GHE, we have conducted two sets of controlled experiments, and they are all based on GTSRB and extracted with HOG. The first set is none-using of GHE in preprocessing, and employing ELM and Kernel ELM as classifier respectively to train, known as HOG+ELM, HOG+KELM. The second set is using GHE, named as GHE+HOG+ELM, and GHE+HOG+KELM.

Table 1: The testing accuracy(%) comparisons of usage of GHE based on ELM and Kernel ELM

Methods	Speed limits	Other prohibitions	Derestriction	Mandatory	Unique	Danger	Average
HOG+ELM	98.97	100	99.28	99.78	99.80	99.25	99.51
GHE+HOG+ELM	99.06	100	99.32	99.84	100	99.74	99.66
HOG+KELM	98.73	100	99.17	99.57	99.80	99.08	99.392
GHE+HOG+KELM	98.97	100	99.17	99.84	100	99.61	99.598

Compared with the results before and after GHE referred as Table 1, the methods with GHE performs better than methods without GHE on each individual category. Also, the Fig 5 represents the visualized effects of GHE.

4.5 Evaluation of fusion of HOG and Gabor features

On account of **NO** faults is tolerable for AVs, as we suggested, HOG and Gabor features should be fused to gain better performance because the fusion of them

would enrich representations of original images, and enhance denoising ability. To testify this, we have compared our STF-ELM proposed fusion of HOG and Gabor as HOG+Gabor-STF with concatenation of HOG and Gabor as HOG+Gabor-CON, and individual feature. The accuracy of each is shown in Table 2. As the table shows, on average, HOG features extracted from datasets outperforms Gabor, and it achieves 100% accuracy on two categories. However, for AVs, none of those mistakes is permitted while driving, so HOG is not enough.

From the table we can also conclude that STF-ELM proposed feature HOG+Gabor-STF, the fusion of HOG and Gabor, achieves the highest average accuracy of 99.83%, and especially, it achieves 100% accuracy on five super categories except just one. The combination of two features, HOG+Gabor-CON, with extremely high dimension of 11176, introduces a lot of noises and superfluous information, resulting in lowest accuracy. So, compared with other existed models, features of STF-ELM is the best for AVs.

Table 2: Recognition accuracy(%) of different features on super categories of GTSRB using ELM as classifier

Methods	Speed limits	Other prohibitions	Derestriction	Mandatory	Unique	Danger	Average
GHE+HOG+ELM	99.06	100	99.32	99.84	100	99.74	99.66
GHE+Gabor+ELM	95.04	99.71	99.58	99.53	100	99.25	98.852
HOG+Gabor-CON	95.51	99.86	87.86	99.57	99.75	99.80	97.058
HOG+Gabor-STF	**98.98**	**100**	**100**	**100**	**100**	**100**	**99.83**
HOGv+r(ELM) [2]	99.14	99.80	96.94	99.77	97.81	99.90	98.893

4.6 Overall recognition performance evaluation

As a whole, models should be tested for accuracy and efficiency, and compared with other published results for overall performance estimation. There are other six proposed TSR models performed on GTSRB from published papers. The description of each model is listed as below, and results of them are compared in Table 3.

1. DP-KELM [22] was implemented on GTSRB transformed to Lab space and used CNN to extract deep perceptual features(DPF), and then used kernel ELM as classifier.
2. Graph LDA [7] combined higher order spectra(HOS) and texture feature as representations of original images and employed linear discriminant analysis(LDA) as classifier.
3. HOGv+r(KELM) based [2] extracted HOG variant as HOGv and use kernel ELM as classifier.
4. Hierarchy SVM [11] designed a hierarchical method for classification by perspective adjustment based on SVM.
5. Ensemble CNNs [15] implemented hinge loss stochastic gradient descent(HLSGD) on CNN training process.

6. Cascaded CNNs [36] designed a two-stage cascaded CNN which utilized attribute-supervisory.

Table 3: Overall accuracy(%) and performance comparison with different proposed models

Models	Speed limits	Other prohibitions	Derestriction	Mandatory	Unique	Danger	Average
STF-ELM	**98.98**	**100**	**100**	**100**	**100**	**100**	**99.83**
STF-KELM	99.11	100	99.67	100	99.84	99.61	99.705
HOGv+r(KELM) [2]	99.54	100	98.33	99.94	99.95	98.96	99.453
HOGv+r(ELM) [2]	99.14	99.80	96.94	99.77	97.81	99.90	98.893
DP-KELM [22]	N/A	N/A	N/A	N/A	N/A	N/A	99.54
Hierarchical SVM [11]	N/A	99.63	98.89	99.94	99.90	99.03	99.478
Cascaded CNNs [36]	99.15	99.01	97.03	94.95	99.85	96.22	97.94
Human(best individual) [37]	98.32	99.87	98.89	100	100	99.21	99.382
Human(average) [37]	97.63	99.93	98.89	99.72	100	98.67	99.14

From Table 3, we can find that our proposed model STF-ELM based on ELM performs best with the highest testing accuracy of 99.83% and achieves 100% on five categories of six. Also, our model based on KELM gets the second place with accuracy of 99.705%. Compared with other listed ELM-based methods such as HOGv+r(KELM) and DP-KELM, we could suppose that our methods outperform them in TSR whole process. In contrast with other state-of-the-art methods based on CNNs, despite the quite small margins of 0.18% between our methods and Ensemble CNNs [15], the computational costs of CNNs could be as much as dozens or even hundreds of times higher than our ELM based models.

To sum up, in terms of accuracy and efficiency, our TSR model achieves highest and almost perfect recognition accuracy and absolutely competitive efficiency, results in highest practical value to be introduced in driver assistant systems of autonomous vehicles.

5 Conclusion

This paper designs a high-precision and efficient TSR model, known as STF-ELM. This model employs GHE on datasets to balance gray levels distribution and enhance contrast, and designs a new representation of traffic signs, feature fusion of HOG and Gabor features, which is invariant to various inference and more abundant based on spatial, time and frequent domains, then uses regularized ELM as classifier to achieve high efficiency. Our STF-ELM model has been tested on GTSRB benchmark, and the experimental results demonstrates that our proposed method has achieved almost perfect average testing accuracy at 99.83% on six super categories, and five of which reached 100% accuracy. Compared with other state-of-the-art CNNs based TSR methods, our ELM based classifier is quite competitive for its perfect balance of training efficiency and

testing accuracy. For those extraordinary performance, in the future, our proposed STF-ELM is quite suitable to be introduced in driver assistant systems of autonomous vehicles.

Acknowledgments

This work has been supported by the FDCT of Macau under MoST-FDCT Joint funding no: (015/2015/AMJ), and University of Macau under the grant no: MYRG2016-00160-FST.

Bibliography

[1] M. Egmont-Petersen, D. de Ridder, and H. Handels, "Image processing with neural networksa review," *Pattern recognition*, vol. 35, no. 10, pp. 2279–2301, 2002.

[2] Z. Huang, Y. Yu, J. Gu, and H. Liu, "An efficient method for traffic sign recognition based on extreme learning machine," *IEEE transactions on cybernetics*, vol. 47, no. 4, pp. 920–933, 2017.

[3] F. Măriuţ, C. Foşalău, M. Avila, and D. Petrişor, "Detection and recognition of traffic signs using gabor filters," in *Telecommunications and Signal Processing (TSP), 2011 34th International Conference on.* IEEE, 2011, pp. 554–558.

[4] M. Z. Abedin, P. Dhar, and K. Deb, "Traffic sign recognition using hybrid features descriptor and artificial neural network classifier," in *Computer and Information Technology (ICCIT), 2016 19th International Conference on.* IEEE, 2016, pp. 457–462.

[5] M. Takaki and H. Fujiyoshi, "Traffic sign recognition using sift features," *IEEJ Transactions on Electronics, Information and Systems*, vol. 129, pp. 824–831, 2009.

[6] M. C. Kus, M. Gokmen, and S. Etaner-Uyar, "Traffic sign recognition using scale invariant feature transform and color classification," in *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on.* IEEE, 2008, pp. 1–6.

[7] A. Gudigar, S. Chokkadi, U. Raghavendra, and U. R. Acharya, "Local texture patterns for traffic sign recognition using higher order spectra," *Pattern Recognition Letters*, 2017.

[8] L. Juan and O. Gwun, "A comparison of sift, pca-sift and surf," *International Journal of Image Processing (IJIP)*, vol. 3, no. 4, pp. 143–152, 2009.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[10] B. Pourebrahimi and J. C. van der Lubbe, "A novel approach for noise reduction in the gabor time-frequency domain." in *VISAPP (2).* Citeseer, 2009, pp. 22–27.

[11] G. Wang, G. Ren, Z. Wu, Y. Zhao, and L. Jiang, "A hierarchical method for traffic sign classification with support vector machines," in *Neural Networks (IJCNN), The 2013 International Joint Conference on.* IEEE, 2013, pp. 1–6.

[12] F. Zaklouta, B. Stanciulescu, and O. Hamdoun, "Traffic sign classification using kd trees and random forests," in *Neural Networks (IJCNN), The 2011 International Joint Conference on.* IEEE, 2011, pp. 2151–2155.

[13] P. Zhang and Z. Yang, "A robust adaboost. rt based ensemble extreme learning machine," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[14] P.-B. Zhang and Z.-X. Yang, "A novel adaboost framework with robust threshold and structural optimization," *IEEE transactions on cybernetics*, 2017.

[15] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 1991–2000, 2014.

[16] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2110–2118.

[17] T. M. Mitchell, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997.

[18] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[19] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A multi-class classification competition," in *IEEE International Joint Conference on Neural Networks*, 2011, pp. 1453–1460.

[20] Z. Huang, Y. Yu, and J. Gu, "A novel method for traffic sign recognition based on extreme learning machine," in *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*. IEEE, 2014, pp. 1451–1456.

[21] X. Yuan, X. Hao, H. Chen, and X. Wei, "Robust traffic sign recognition based on color global and local oriented edge magnitude patterns," *IEEE transactions on intelligent transportation systems*, vol. 15, no. 4, pp. 1466–1477, 2014.

[22] Y. Zeng, X. Xu, D. Shen, Y. Fang, and Z. Xiao, "Traffic sign recognition using kernel extreme learning machines with deep perceptual features," *IEEE Transactions on Intelligent Transportation Systems*, 2016.

[23] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 2809–2813.

[24] R. Maini and H. Aggarwal, "A comprehensive review of image enhancement techniques," *arXiv preprint arXiv:1003.4053*, 2010.

[25] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 1918–1921.

[26] Z. Huang, Y. Yu, J. Gu, and H. Liu, "An efficient method for traffic sign recognition based on extreme learning machine," *IEEE transactions on cybernetics*, 2016.

[27] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2012.

[28] G.-B. Huang, Z. Bai, L. L. C. Kasun, and C. M. Vong, "Local receptive fields based extreme learning machine," *IEEE Computational Intelligence Magazine*, vol. 10, no. 2, pp. 18–29, 2015.

[29] C. M. Wong, C. M. Vong, P. K. Wong, and J. Cao, "Kernel-based multilayer extreme learning machines for representation learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[30] P. K. Wong, J. Zhong, Z. Yang, and C. M. Vong, "Sparse bayesian extreme learning committee machine for engine simultaneous fault diagnosis," *Neurocomputing*, vol. 174, pp. 331–343, 2016.

[31] V. Kyrki, J.-K. Kamarainen, and H. Kälviäinen, "Simple gabor feature space for invariant object recognition," *Pattern recognition letters*, vol. 25, no. 3, pp. 311–318, 2004.

[32] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognition*, vol. 38, no. 12, pp. 2437–2448, 2005.

[33] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE signal processing letters*, vol. 9, no. 2, pp. 40–42, 2002.

[34] J. Shlens, "A tutorial on principal component analysis," *arXiv preprint arXiv:1404.1100*, 2014.

[35] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 809–821, 2016.

[36] K. Xie, S. Ge, Q. Ye, and Z. Luo, "Traffic sign recognition based on attribute-refinement cascaded convolutional neural networks," in *Pacific Rim Conference on Multimedia*. Springer, 2016, pp. 201–210.

[37] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol. 32, pp. 323–332, 2012.