Cognition based Recognition of Partially Occluded Traffic Signs

Abdul Mannan\textsuperscript{*a,b,1}, Kashif Javed\textsuperscript{b}, Ata ur Rehman\textsuperscript{b}, Haroon A Babri\textsuperscript{b}, Serosh K Noon\textsuperscript{a}

\textsuperscript{a}Department of Electrical Engineering, NFC Institute of Engineering and Technology, Multan, Pakistan-59060
\textsuperscript{b}Department of Electrical Engineering, University of Engineering and Technology, Lahore, Pakistan

Abstract

Computer vision based traffic sign detection and recognition is an active field of research but the task becomes challenging when the sign of interest is partially occluded by nearby objects like a tree, pole or vehicle. Another difficulty posed especially in the developing countries is the lost colors problem that arises due to aging and poor maintenance. This work presents an automatic technique that focuses on visible parts only and suppresses occluded portions. Features are collected using a convolutional neural network inspired invariant feature extraction technique augmented with feature interaction based dimensionality reduction. Further, with the use of dynamic parameter estimation, an adaptive system for continuous learning is also proposed. Since the effect of partial occlusion has not been thoroughly studied, there is no benchmark database available for this purpose. We have prepared two datasets by combining originally and synthetically occluded images taken from field surveys and from famous GTSRB database. Experiments revealed that our technique outperforms state of the art recognition methods previously used for visible and occluded signs by obtaining 0.81 precision and 0.79 recall values on the average. The proposed method also shows a remarkably low error rate as the amount of occlusion is increased.

Keywords: Traffic sign, cognition, occlusion, parameter estimation, discrete cosine transform, dimensionality reduction, feature extraction, convolutional neural networks

1. Introduction

Traffic signs are used worldwide to guide road users with an objective to avoid accidents [1]. Mainly, there are two types of traffic signs i.e. (1) American with white background and black foreground and (2) European having either red rim or filled with blue color [2, 3, 4, 5]. In this work, we focus on European traffic signs only. Many computer vision based methods to automatically detect and recognize European traffic signs have been reported in literature [1, 6, 7, 8, 9, 10]. Detection serves the purpose of segmenting a traffic sign in a real world scene whereas recognition deals with reading its contents. Many automatic traffic sign detection and recognition systems can effectively detect and recognize visible signs in many parts of the world [11]. For detection, various color, saliency and statistical distribution of pixels based methods have been reported in the literature [12, 13, 1]. The next subproblem i.e., recognition has been tackled by researchers by using different techniques to extract invariant features from traffic sign shapes. Some famous methods are histogram of oriented gradients (HOG) [14], local binary patterns (LBP) [15], their combinations [1], transform based methods [1] (e.g., Fourier, discrete wavelet and discrete cosine etc.) and higher order spectra [16]. Feature extraction can be applied on an image as a whole or with the help of a rectangular or polar grid [2]. Reducing the size of feature vector by getting rid of irrelevant and redundant features, dimensionality reduction [17] has been proved to be an effective method.

\textsuperscript{1}Corresponding Author
Email: mannan@nfciet.edu.pk
Phone: 00923216301074

Preprint submitted to Elsevier February 8, 2022
Dimensionality reduction not only enhances computational speed but also increases the recognition accuracy [18]. A multiclass classifier is then employed to assign traffic signs their correct classes [19]. An alternate approach to recognizing contents of a traffic sign is using deep neural networks [20, 21] which embed both feature learning and classification stages. However, internal architecture of such networks is complex and requires learning of thousands (or even millions) of parameters. On small datasets containing only a few thousand images, these networks are prone to over-fitting [22]. Pre-trained models of famous convolutional neural networks like Alexnet or Googlenet [23, 24] can, however, be customized to fit for a small dataset [25].

In some cases, a traffic sign may be occluded by other nearby objects [26, 27]. This occlusion can be caused by a tree, vehicle, pole or can be deliberate e.g., by pasting a sticker or writing something on the sign board. Detecting and recognizing an occluded traffic sign is challenging because some portion of the object of interest is not visible. For partially occluded traffic sign images, extracting meaningful features is a bit tricky because we have to take care of the occluded portion too. We find some efforts for detection of occluded traffic signs in the literature including the novel pre-processing layer based solution proposed by us [28] but there is indeed a little work on the recognition of their contents.

We have identified following four categories of occlusions commonly found for traffic signs:

- Natural
- Incidental
- Deliberate
- Degradation

Natural occlusion is caused by a nearby tree branch or leaves, incidental is the occlusion by chance due to the presence of a vehicle or pole. The third category of occlusion occurs when humans deliberately suppress the contents of a traffic sign by putting an advertisement sticker on it. In certain parts of the world where the sunshine is bright and/or due to lack of adequate maintenance, traffic signs are frequently found degraded (i.e., category 4). Examples of each of the four types of occlusions are shown in Figure 1.

We, in this work, address the problem of recognizing contents of occluded and degraded traffic signs by taking inspiration from human cognition. Instead of working in RGB domain, we obtain a boosted image with the help of linear mapping and location aware edge filtering. In this image, portions related to the occlusions are suppressed and the visible parts are highlighted so that our convolutional neural network inspired multiscale spectral feature extraction method can concentrate on visible portions only. A dimensionality reduction technique using feature interaction has also been employed to reduce the size of data which results in lesser computational cost and higher recognition accuracy. Finally, a multiclass support vector machine (SVM) classifier is used to make predictions. In order to address the subproblem of degradation, a dynamic parameters estimation technique with feedback is also proposed.

The remainder of this paper is organized as follows: Section 2 describes some related work especially pertaining to occluded and degraded traffic sign’s detection and recognition, Section 3 describes our proposed technique to recognize occluded and degraded traffic signs. Section 4 explains experimental setup and Section 5 describes results obtained using different settings. In Section 6, we discuss experimental results of our proposed method and finally Section 7 describes conclusion of the work.

2. Related Work

A detailed review of standard and visible traffic sign detection and recognition techniques can be found in [1] and [29]. We, in this section, will limit our discussion to detection and recognition
of occluded traffic signs only.

Though automatic traffic sign detection and recognition is an active field of research, dealing with occluded and degraded signs has been least addressed. We find only a few efforts to handle this problem in literature. Occlusion maps based technique in [27] uses decomposition of support vector machine (SVM) score to detect possible traffic signs in a given real world scene but the work is limited to the detection and super class identification i.e. circle, triangle etc. Occlusion maps were originally proposed by Wang et al [30] to detect occluded objects in images; they use a combination of histogram of oriented gradients (HOG) and local binary patterns (LBP) to detect occluded objects of interest. Rehman et al.in [31] used a scheme to automatically identify occluded portions during training process by using a predefined heuristic. Then the features were collected of visible portions only called discriminant patches (d-patches). A disadvantage of this approach is that the performance degrades severely with increasing amount of occlusion on test data. A similar work for detecting traffic signs in real world scenes is presented in [15]. They proposed the use of a large number of color channels from various color spaces and applied local binary patterns (LBP) as feature extraction technique. Finally, integral features were used to generate a large feature vector. Floros et al. [32] developed a scheme to detect degraded traffic signs in natural scenes using RGB thresholding and heuristic based search. For experimentation, they collected a small dataset of around three hundred images divided in only five classes from two different places in Greece. Li et al. [33] used a faster version of probabilistic neural networks to recognize occluded and degraded traffic signs on a synthetic dataset of few hundred images but is prone to over fitting due to small training and test datasets. A generative model was trained and used by Ishida et al. [34] to identify degraded traffic signs taken from a synthetically generated dataset of standard and degraded traffic signs. All methods mentioned above are mainly used to detect partially occluded and/or degraded traffic signs in real world scenes leaving recognition of contents as a future work.

To recognize contents of partially occluded traffic signs, some relatively older systems designed for standard and visible signs [35, 26] also respond well. Fleyeh et al., [26] made use of eigen space and consider only top few eigen vectors as features; these were found effective for some partial
occlusions too. Escalera et al., [35] used deformable models based on energy of color and gradient to detect traffic sign in natural scenes but they used a very small dataset for evaluation.

Keeping in view partial occlusions, we find a number of techniques to identify objects in general, but there are only a few methods pertaining to traffic signs’ detection and recognition.

3. Our Proposed Technique

In this section, we describe at length, our proposed method for recognition of partially occluded traffic signs. Block diagram of the proposed system is shown in Figure 2 and a brief description of different blocks is given in the following sections.

3.1. Linear Transformation

The objective of linear transformation is to map RGB input data to another color space where:

- pixels corresponding to different colors map far from each other and
- pixels corresponding to the same color map very close to each other.

As shown in the top left corner of Figure 2, a data matrix \( A \) is created by putting RGB data corresponding to red and blue colors as its rows. The data in each row is then transformed to an output matrix \( B \) according to the linear transformation mentioned in Equation 1.

\[
b(j) = T_{lin}[a(j)] = a(j) \times E \tag{1}\]

At any arbitrary row \( j \) in the data matrix \( A \), \( a(j) = [a_1(j) \ a_2(j) \ a_3(j)] \) is a tuple containing R, G and B color channels whereas \( b(j) = [b_1(j) \ b_2(j)] \) is the data at the same row in the transformed matrix \( B \). \( a(j) \), a \( 1 \times 3 \) vector is multiplied by the transformation matrix \( E \) of size \( 3 \times 2 \) to give \( b(j) \) of size \( 1 \times 2 \).

3.1.1. Computing the Coefficient of Matrix \( E \)

To compute the coefficients of matrix \( E \), the RGB pixel data contained in matrix \( A \) is mapped on the Eigen space and the two vectors pertaining to the largest two Eigen values are retained [36, 37].

The process includes the following steps:

- Covariance matrix \( \Sigma^A \) (of size \( 3 \times 3 \)) was computed for the data matrix \( A \) (of size \( n \times 3 \)), where \( n \) is the number of rows in matrix \( A \)
- Three Eigen values \( \lambda_1, \lambda_2 \) and \( \lambda_3 \) were computed for the covariance matrix \( \Sigma^A \) such that \( \lambda_1 > \lambda_2 > \lambda_3 \)
- Two Eigen vectors \( \mathbf{w}_1 \) and \( \mathbf{w}_2 \) corresponding to the two largest Eigen values \( \lambda_1 \) and \( \lambda_2 \) were computed and were put as the columns of the matrix \( E \) i.e., \( E = [\mathbf{w}_1 \ \mathbf{w}_2] \). This gives the matrix \( E \) of size \( 3 \times 2 \).

This linear transformation tries to maximize separation among colors pertaining to the traffic signs and those taken by the occlusions and the background in addition to the advantage of reducing input data dimensions from three to two [38].
Figure 2: Block diagram of the proposed system.
3.2. Dynamic Parameter Estimation

As a result of applying Equation 1 on each row of 3-dimensional data matrix $A$, a 2-dimensional transformed matrix $B$ is obtained. The next step is to estimate parameters of the two color classes $(k)$, assuming each of them to be normally distributed i.e.,

$$P(b|c_k) \sim \mathcal{N}(b; \mu_k, \Sigma_k), \quad k = 1, 2$$

(2)

Maximum likelihood estimates [39] of multivariate mean $\mu_k$ and covariance $\Sigma_k$ of training data are given in Equations 3 and 4. It is assumed that there are $R$ training images having $P$ pixels of the color $k$ per training image. As mentioned earlier, $b(j)$ is the two dimensional data present at location $j$ in data matrix $B$. To be more precise, $b_{r,p,k}(j)$ is a tuple of two dimensional data corresponding to $p^{th}$ pixel in $r^{th}$ training image pertaining to class $k$ present at any arbitrary location $j$ in data matrix $B$. Since index $j$ is obvious, we will drop it for subsequent discussions.

$$\mu^t_k = \frac{\sum_{r=1}^{R} \sum_{p=1}^{P} b^t_{r,p,k}}{R \times P} \quad (3)$$

$$\Sigma^t_k = \frac{\sum_{r=1}^{R} \sum_{p=1}^{P} (b^t_{r,p,k} - \mu^t_k)(b^t_{r,p,k} - \mu^t_k)^T}{R \times P} \quad (4)$$

3.3. Boosted Image

Given an image $G$ in transformed color space obtained by applying Equation 1 on RGB image, likelihood function mentioned in Equation 2 is used to know how likely each pixel in the transformed domain image $G$ belongs to class $c_k$. With this information available, posterior probability is computed using Bayes rule mentioned in Equation 5. $P(g(x,y)|c_k)$ is the likelihood that pixel $g(x,y)$ belongs to class $c_k$, $P(g(x,y))$ is evidence and $P(c_k)$ is the prior probability. The prior depends on the proportional representation of samples from each class and is computed using Equation 6. Here $N_k$ is the number of instances in training data belonging to class $c_k$ and $N_T$ are the total number of training instances.

$$P(c_k|g(x,y)) = \frac{P(g(x,y)|c_k)P(c_k)}{P(g(x,y))} \quad (5)$$

$$P(c_k) = \frac{N_k}{N_T} \quad (6)$$

Posterior probability $P(c_k|g(x,y))$ corresponds to a grayscale image in which pixels corresponding to the colors of interest i.e., red and blue are intended to have significantly higher grayvalues than other pixels. Red and blue colors are chosen because the European traffic signs we are interested in appear either with a red outer rim or are filled with blue color [40]. In order to highlight the contents which are either black foreground on white background or white foreground on blue background, location aware edge filtering $edge^h_l[g(x,y)]$ is employed. Vector $l$ is a set of spatial locations to perform filtering while $h$ is the minimum edge strength to be retained. Values of $l$ and $h$ were empirically computed from the training data. Using this technique, high contrast edges, present only inside the traffic sign rim i.e., middle of the shape are highlighted whereas all other edges are suppressed. As a result, the pixels corresponding to the contents of the sign take higher gray values than the rest. It is to be noted that any occlusions present in the middle of the shape having weak edges is also suppressed.
Figure 3: Graphical description of the proposed location-aware edge filtering by varying parameters $h$ and $I$. The filtered image obtained using $I_{\text{High}}$ and $h_{\text{Medium}}$ produced the best results.

The operation $\text{edge}^h_I[g(x,y)]$ is further explained with the help of a graphical example shown in Figure 3 for different values of $h$ and $I$. The subscript “Low” indicates that even very weak edges were captured whereas the subscript “High” indicates that the only very strong edges were retained. “Medium” is the edge strength in between the former two options. It can be clearly seen that the best output of the location-aware edge filtering can be ensured if even very strong edges in the background are suppressed and the medium edge strength in the foreground is retained (refer to Figure 3(e)).

Finally, the sum of data corresponding to the posterior i.e., $P(c_k|g(x,y))$ and location aware edge filtering $\text{edge}^h_I[g(x,y)]$ is linearly transformed to form the gray values of the desired boosted image $I$, mathematically

$$i(x,y) = 255 \times P(c_k|g(x,y)) + \text{edge}^h_I[g(x,y)]$$

where $i(x,y)$ is a scalar number equal to the grayvalue of the image $I$ at location $(x,y)$.

In the boosted image $I$ the pixels representing the traffic sign take higher gray levels than the rest. An example is shown in Figure 4 where a sample occluded image shown in Figure 4(a) was processed to obtain the boosted image $I$ shown in Figure 4(b). Further, Figure 4(c) shows surface plot of the boosted image showing that pixels corresponding to the traffic sign take on higher grayvalues and the effect of occlusion and background is suppressed.

3.4. Multiscale Spectral Feature Extraction

The ability of discrete cosine transform (DCT) for feature extraction in images has already been explored [41, 42, 2, 43]. Ayyalasomayajula et al. [42] proposed DCT phase (i.e., sign, either positive or negative) based descriptor to recognize partially occluded real world images. To use for partially occluded and degraded traffic sign recognition, we extend their method as follows:
In order to capture detailed features, we, unlike [42], apply feature extraction method at multiple spatial scales on boosted image $I$ instead of the original RGB image. In Section 3.3, we have mentioned that in the boosted image, the effect of occlusion has already been partially suppressed, which makes the task of feature extraction technique easier.

Since the resulting descriptor is high dimensional, Ayyalasomayajula et al. [42] used empirically determined thresholds to select important features. We, on the other hand, use a more robust technique; feature interaction based dimensionality reduction is employed to select the most relevant and the least redundant information.

Given a boosted spatial domain image $I$, let’s represent the discrete cosine transform of a pixel $i(x, y)$ as $i(u, v)$ by using indices $(x, y)$ for spatial and $(u, v)$ for frequency domain. Mathematical expression to compute the transform of an image of size $X \times Y$ is given in Equation 8 where $\alpha_u$ and $\alpha_v$ are simple functions of indices $u$ and $v$ [44]. The real valued pixel in transformed domain i.e., $i(u, v)$ can be written in terms of magnitude component which is its absolute value $i_{abs}(u, v)$ and the phase term that is the sign either positive or negative $i_{ph}(u, v)$ as stated in Equation 9.

$$i(u, v) = T_{dct}[i(x, y)] = \alpha_u \alpha_v \sum_{x=0}^{X-1} \sum_{y=0}^{Y-1} i(x, y) \cos \frac{\pi(2x+1)u}{2X} \cos \frac{\pi(2y+1)v}{2Y}$$

$$0 \leq u \leq X - 1 \quad 0 \leq v \leq Y - 1$$

$$i(u, v) = i_{abs}(u, v) \times i_{ph}(u, v)$$

where

$$i_{abs}(u, v) = \|i(u, v)\|$$
Figure 5: An illustration of discrete cosine transformed magnitude and phase reconstruction.

\[ i_{ph}(u,v) = \frac{i(u,v)}{i_{abs}(u,v)} \]  \hspace{1cm} (11)

and hence

\[ I = I_{abs} \circ I_{ph} \]  \hspace{1cm} (12)

where “\( \circ \)” refers to the Hadamard product operation [45] where element wise multiplication is performed between the two matrices \( I_{abs} \) (containing only the magnitude) and \( I_{ph} \) (containing only the phase).

Figure 5 shows a sample input image containing a traffic sign and the images reconstructed from the magnitude and phase of its discrete cosine transform. It can be seen that the phase reconstructed image [46] contains more visual information compared to magnitude reconstructed image. This renders the DCT transformed phase image more suitable for the purpose of extracting invariant features [42].

In order to obtain an occlusion invariant descriptor, we propose a feature extraction strategy inspired by famous convolutional neural networks [12]. Block diagram of the feature extraction technique is shown in Figure 6. Each input image in the training set is passed through two average pooling layers where average operation is performed on a box of \( 2 \times 2 \) pixels which in turn reduces the size of the image to one half for the subsequent layer (i.e., filter size = \( 2 \times 2 \) and stride = \( 2f \)). DCT phase features are computed on images tapped at different intermediate points (i.e., scales) and the features are finally concatenated. It should be noted that at each layer, feature extraction operation is performed on blocks of size \( 8 \times 8 \) which was empirically found to be the best choice.

Mathematical representation of this operation on a sample image \( I_i \) taken from the training dataset is shown in Equation 13.

\[
d_i = \text{cat}\{T_{dct,ph}^{64\times64}[I_i] \ T_{dct,ph}^{32\times32}[I_i] \ T_{dct,ph}^{16\times16}[I_i]\} \hspace{1cm} (13)
\]

As shown in Figure 6, \( d_i \) is a long (4096 + 1024 + 256 = 5376 dimensional) descriptor for each training image \( I_i \). Finally, we come up with a data matrix \( D = [d_1 \ d_2 \ d_3 \ \ldots \ d_R]^T \) where each row contains a vector descriptor \( d_i \) for exactly one training image. In order to extract only the relevant and least redundant features of this high dimensional feature matrix \( D \), we use a feature interaction based dimensionality reduction strategy described in Section 3.5.

### 3.4.1. Obtaining Invariance to Occlusions

To show that the proposed multiscale feature extraction technique is invariant to occlusions, let’s pick a small subimage (called \( I_{new}^{sub} \)) from the boosted image \( I \) containing some occlusion.
Irrespective of the noise source (but assuming linearity), $I_{\text{new}}$ can mathematically be decomposed into the original visible subimage $I_{\text{sub}}$ plus a noisy subimage $I_{\text{noise}}$, i.e.,

$$I_{\text{new}} = I_{\text{sub}} + I_{\text{noise}}$$  \hspace{1cm} (14)

A pictorial representation of the situation is given in Figure 7. At multiscale level $n \times n$, the descriptor computed using our proposed method for subimage $I_{\text{new}}$ at position $(x, y)$ can be extracted as [47]:

$$\text{sign}(T_{\text{dctph}}^{n \times n}[i_{\text{new}}(x, y)]) = \text{sign}\{T_{\text{dctph}}^{n \times n}[i_{\text{sub}}(x, y)] + T_{\text{dctph}}^{n \times n}[i_{\text{noise}}(x, y)]\}$$  \hspace{1cm} (15)

The operation is distributive because $T_{\text{dctph}}^{n \times n}$ is a linear operator. For simplicity, $i_{\text{noise}}(x, y)$ a pixel from noisy image ($I_{\text{noise}}$) is considered to be a random variable sampled from a one dimensional Gaussian distribution with mean ($\mu_0$) and variance ($\sigma_0^2$), i.e.,

$$i_{\text{noise}}(x, y) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$  \hspace{1cm} (16)

The parameters are estimated as follows:

$$\mu_0 = E[i_{\text{noise}}(x, y)]$$  \hspace{1cm} (17)

$$\sigma_0^2 = E[(i_{\text{noise}}(x, y) - \mu_o)^2]$$  \hspace{1cm} (18)

As a result, an occluded or degraded pixel in image $I_{\text{sub}}^{\text{new}}$ can be written as:

$$i_{\text{new}}^{\text{sub}}(x, y) = i_{\text{sub}}(x, y) \pm K \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(i_{\text{noise}}^{\text{sub}}(x, y) - \mu_0)^2}{2\sigma_0^2}}$$  \hspace{1cm} (19)

Here, $K$ is an empirically chosen constant that determines the weight given to the random variable drawn from the Gaussian distribution $\mathcal{N}(\mu_0, \sigma_0^2)$. The scheme given in Equation 19 can be used to simulate occlusions on a given sample traffic sign image. An example is given in Figure 8 where an occlusion is simulated on the top left corner of a SL30 traffic sign. Figure 8(a) gives a scenario where an improper value of $K$ causes a greenish blur on the targeted area failing to introduce a significant effect of occlusion. In contrast, in Figure 8(b) a black patch over the desired area is visible which properly simulates the effect of occlusion.
Moreover, for any two scalars \( w \) and \( x \), the sign of the sum is given by [48]:

\[
\text{sign}(w + x) = \text{sign}(w) \quad \text{if} \quad |w| > |x|
\]  

(20)

Since \( T_{\text{dct,ph}}^{n \times n} [i_{\text{noise}}(x, y)] \) is computed using a sparse noisy image and obtained as a result of small positive or negative addition in the original subimage (refer to Figure 7), its magnitude is small compared to \( T_{\text{dct,ph}}^{n \times n} [i_{\text{sub}}(x, y)] \). Therefore, right hand side of Equation 15 tends to retain the sign of \( T_{\text{dct,ph}}^{n \times n} [i_{\text{sub}}(x, y)] \) and hence our proposed method remains stable in the presence of occlusions.

3.5. Feature Interaction based Dimensionality Reduction

Since some information in the long feature vector matrix \( D \) is more useful than the other, there is a need to select features that are strongly correlated with the target class (i.e., the contents of the sign e.g., SL 100, Right Turn etc.) and are least redundant with each other [49]. The operation is shown with the help “Finding useful features” block in Figure 2. The high dimensional feature matrix \( D \) is fed to the dimensionality reduction operation and the output is a lower dimensional matrix \( S \) of the most important features. As mentioned in Section 3.4, discrete cosine transform (DCT) phase information is the sign of the transform which is represented as a binary data i.e., either +1 or -1. This facilitates the use of mutual information based dimensionality reduction technique.

To obtain a subset of the most relevant and least redundant features, we propose the use of feature interaction based dimensionality reduction [50, 51]. Assuming columns of the data matrix \( D \) are represented with \( f \), main steps of the dimensionality reduction technique are given as under:
1. $MI(f; c)$ is mutual information based relevance between a class $c$ and a feature $f$. For all features in the dataset $D$, this procedure is repeated and finally the features put in descending order with respect to their relevance with the target class.

2. The top ranked feature is the one that scores the highest class relevance. This feature is directly put into the subset of selected features $S$ without any comparison.

3. The second highest feature with respect to the class relevance ($f^+$) is temporarily put in $S$ and mutual information between this new subset $\{f^+, S\}$ and the class $c$ is computed i.e., $MI(f^+, S; c)$.

4. Step 3 is repeated for all remaining features in $D$ and the feature with the highest value of $MI(f^+, S; c)$ is added to the subset $S$ permanently.

5. In order to obtain required size of subset $S$, the processes in steps 3 and 4 are repeated equal to the number of required features.

An exact formula to compute joint mutual information (in step 4) between candidate feature ($f^+$) temporarily added to the already selected subset of features ($S$) and the target class ($c$) is given by Equation 21. Right hand side of the equation shows the same expression but with the subset ($S$) written in terms of its columns ($s_1, s_2, s_3, ..., s_s$). Each column of the subset $S$ expresses one feature.

$$MI(f^+, S; c) = MI(f^+, s_1, s_2, s_3,...s_s; c) \tag{21}$$

Since complex joint probability terms are involved in its direct computation which makes it computationally expensive, it’s approximation given in Equation 22 is used i.e., features are taken from $S$ one at a time, joint mutual information is computed and the results are finally summed together [52].

$$MI(f^+, S; c) \approx \sum_{s_s \in S} MI(f^+, s_s; c) \tag{22}$$

Each term on the right hand side of Equation 22 can be expanded and written in the form of its joint and marginal entropy equivalents as given by Equation 23. $H(f^+, f_s)$ is the joint entropy term between the temporarily added feature and a sample feature picked from $S$, $H(c)$ is the entropy of class label and $H(f^+, s_s, c)$ is another joint entropy term.

$$MI(f^+, s_s; c) = H(f^+, s_s) + H(c) - H(f^+, s_s, c) \tag{23}$$

It is noteworthy that since $H(c)$, the entropy of class label, remains the same throughout the feature selection procedure, therefore it can be dropped and the new computational formula for approximate joint mutual information ($MI'$) is given in Equation 24. This approximation was used on our occlusion invariant multiscale feature data matrix $D$ to obain a subset of the most important features $S$.

$$MI'(f^+, s_s; c) = H(f^+, s_s) - H(f^+, s_s, c) \tag{24}$$

4. Experiments

We tested our proposed method for recognition of occluded traffic signs on two datasets namely Dataset1 and Dataset2. Detailed description of each dataset will be given in next section. After feature extraction and dimensionality reduction steps, a classifier was trained separately for both datasets. To compare our proposed method with other state of the art and recent methods, various evaluation measures like precision, recall and recognition accuracy were used. Performance enhancement in terms of computational complexity and execution time was also monitored for the dimensionality reduction algorithm employed by our proposed technique.
4.1. **Dataset**

There is no publicly available dataset of occluded and degraded traffic signs right now. We collected around 1,000 real world images containing partially occluded or degraded traffic signs by traveling on N5 and M2 highways in Pakistan [53]. To complete the dataset, a large number of images were synthetically produced using commonly found occlusions i.e., tree leaves, vehicle, buildings and other installations. Dataset1 contains around 1,903 images divided in 18 classes. A second dataset (Dataset2) is composed of 1,321 images in 26 classes and was compiled using naturally occluded samples taken from GTSRB dataset [54] and synthetically occluded or degraded examples. Table 1 contains detailed information about the datasets used for experimentation. All images were rescaled to 64 × 64 to facilitate multiscale spectral feature extraction. Figure 9 shows some samples from the two datasets used for experiments. This is a maiden work to use real world occlusions for experiments, all previous attempts for occluded and degraded traffic signs have used synthetic images only.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Instances</th>
<th>Classes</th>
<th>Class Names</th>
<th>Composition</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset1</td>
<td>1,903</td>
<td>18</td>
<td>SL20, SL30, SL50, SL70, SL90, SL100, SL120, Slow, Warning right, Warning left, Round about, Falling Rocks, School, Traffic signals, Road works, Road narrow, Roads meeting, Pedestrian</td>
<td>Originally occluded, Originally degraded, Synthetically occluded, Standard, Negative</td>
<td>Field survey, Field survey, Occlusions from field data, Field survey</td>
</tr>
<tr>
<td>Dataset2</td>
<td>1,321</td>
<td>26</td>
<td>SL20, SL30, SL50, SL60, SL70, SL80, SL100, Proceed straight, Turn left, Turn right, Pass on left, Pass on right, Round about, Warning right, Warning left, Danger point, Traffic signals, Road works, Road narrow, Pedestrian, Bicycle, Kids crossing, Speed bump, Slippery road, Road bend, Roads meeting</td>
<td>Originally occluded, Synthetically occluded, Standard, Negative</td>
<td>GTSRB, Occlusions from field data, GTSRB, Field survey &amp; GTSRB</td>
</tr>
</tbody>
</table>
4.2. Classifiers

To evaluate our proposed method on two datasets (i.e. Dataset1 and Dataset2), a multiclass support vector machine classifier [19] was trained on various reduced subsets of relevant and unique features present in data matrix $S$ first on 18 classes for Dataset1 and then for 26 classes on Dataset2. Details of the datasets are given in Table 1. We used LibSVM [55] implementation of SVM classifier with linear kernel and other default parameters for experimentation. The classifier module is written in C language and supports classification among multiple classes. A core i7 computer with 8 GB RAM was used for experimentation.

4.3. Evaluation Procedure

To evaluate the effectiveness of our proposed method, the available data was divided such that 60% data was used for training and the remaining 40% was kept aside for testing purposes. In order to reduce the chance of over fitting, 5 random combinations of this split were tried and the results were finally averaged.

4.4. Evaluation Metrics

For experimental results on two datasets using our proposed techniques, following evaluation metrics were used [39]:

$$\text{Precision} = \frac{TP}{TP + FP}$$  (25)

$$\text{Recall} = \frac{TP}{TP + FN}$$  (26)

$$\text{Recognition Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$  (27)

$$\text{Error rate} = \frac{FP}{FP + TN}$$  (28)

where TP: true positive, FP: false positive, TN: true negative and FN: false negative.

4.5. Statistical Evaluation Procedure

Block diagram of the proposed evaluation system is given in Figure 10. The input is a test RGB image that is first converted to two channel transformed space by applying Equation 1 on every pixel. The parameters of conversion (contained in matrix $E$) are available after the training is over. Boosted image is obtained next by applying dynamic parameters of distributions i.e., $\mu_i$ and $\Sigma_i$ which also are the result of training. The boosted image is then fed to the feature extraction followed by the dimensionality reduction blocks. The parameters of dimensionality reduction i.e., the set containing information of the most relevant and least redundant features is available as a result of training too. After number of images equal to 5% of the training set have been recognized to contain a traffic sign, RGB information of the two color classes is taken as a sample (called $A_{aug}$) and a statistical test is performed by considering the current RGB training data as another sample. Following null and alternate hypotheses [56] were tested at 95% confidence interval:

$$H_0 : \mu_A = \mu_{A_{aug}}$$

$$H_1 : \mu_A \neq \mu_{A_{aug}}$$

If the two sample means are statistically the same, the new sample of observations is augmented to the $A$ matrix shown in Figure 2. For this purpose the following Hotelling’s $t^2$ statistic [57] is computed:

$$t^2 = \frac{n_A n_{A_{aug}}}{n_A + n_{A_{aug}}} (\mu_A - \mu_{A_{aug}})^T \Sigma_p^{-1} (\mu_A - \mu_{A_{aug}})$$  (29)
Figure 10: Block diagram of the proposed system for testing and continuous updates.
Here, \( n_A \) and \( n_{A_{aug}} \) are the total number of samples in sets \( A \) and in \( A_{aug} \) respectively and \( \Sigma_p \) is pooled covariance computed from the covariance matrices of sets \( A \) and in \( A_{aug} \) using the following formula:

\[
\Sigma_p = (n_A - 1) \Sigma_A + (n_{A_{aug}} - 1) \Sigma_{A_{aug}}
\]  

(30)

The above hypothesis is tested at 95% confidence interval and if the two means \( \mu_A \) and \( \mu_{A_{aug}} \) are found to be statistically equal, the new \( A \) matrix becomes:

\[
A_{t=t_1} = [A_{t=t_0}^T \ A_{aug}^T]^T
\]  

(31)

The system keeps on training itself dynamically and new estimates of mean and covariance are available after a number of test images are passed through the recognition step. This makes the system adaptive to new conditions and is able to continuously update itself in a newly added place anywhere in the world. Consequently, after a number of samples with degraded images are introduced, the mean and covariance estimates are adjusted automatically to accommodate the new situation. A graphical illustration of the concept is shown in Figure 11, the system is trained on a set of standard traffic signs at time \( t = t_0 \) initially and then the parameters keep on updating as it runs on test data. For the sake of understanding it better, a one dimensional representation of the system is shown in Figure 12. The blue curve (drawn as a continuous line) shows parameter estimates corresponding to the RGB data (\( A \)) available in the beginning i.e., at \( t = t_0 \). The red dotted curve shows how it is updated at \( t = t_1 \) as new data (\( A_{aug} \)) corresponding to the degraded signs is added at run time. This addition when augmented to the original data, shifts both mean and variance estimates. Since the new data matrix \( A_{t=t_1} \) has larger size compared to the matrix \( A_{t=t_0} \), only last \( z \) entries in \( A_{t=t_1} \) matrix are used to update parameters where \( z = \text{size}(A_{t=t_0}) \).
5. Results

5.1. Comparing Feature Extraction Techniques

In this section, we compare a number of feature extraction techniques with our proposed multiscale phase spectrum based method. Considering precision and recall as measures, Table 2 shows comparison of our proposed method with recent and state of the art techniques previously used for recognition of visible and/or occluded traffic signs. A recent method based on occlusion maps presented by Hou et al. [27] and two other methods i.e., Eigen [26] and energy [35] based methods were found partially useful for the recognition of occluded and degraded traffic signs. However, state of the art feature extraction techniques such as histogram of oriented gradients (HOG), local binary patterns (LBP) and their combination perform poorly for occluded traffic signs. However, our proposed technique obtains the highest precision and recall for both datasets.

Table 2: Comparison of “Precision” and “Recall” among different methods. In literature, HOG and LBP were used for visible signs only whereas Occlusion maps, Eigen and Energy based methods were found partially useful for occluded traffic signs.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset1</td>
<td>0.61</td>
<td>0.60</td>
<td>0.65</td>
<td>0.67</td>
<td>0.59</td>
<td>0.51</td>
<td>0.81</td>
</tr>
<tr>
<td>Dataset2</td>
<td>0.63</td>
<td>0.59</td>
<td>0.52</td>
<td>0.49</td>
<td>0.78</td>
<td>0.66</td>
<td>0.79</td>
</tr>
<tr>
<td>Recall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataset1</td>
<td>0.59</td>
<td>0.48</td>
<td>0.65</td>
<td>0.61</td>
<td>0.66</td>
<td>0.65</td>
<td>0.79</td>
</tr>
<tr>
<td>Dataset2</td>
<td>0.43</td>
<td>0.49</td>
<td>0.62</td>
<td>0.59</td>
<td>0.68</td>
<td>0.71</td>
<td>0.76</td>
</tr>
</tbody>
</table>

As mentioned in Section 1, we divided the commonly found occlusions in four categories namely Natural, Incidental, Deliberate and Degradations. Table 3 shows recognition accuracy of these four categories of occlusions mentioned versus recent and state of the art methods for both datasets.
Our proposed method outperformed the other four methods by being the most accurate on both datasets. Moreover, the method was found to be the most effective for all categories of occlusions. Eigen Space [26] and Deformable Models [35] were found better than occlusion maps based technique [27]. This is because occlusion maps based method [27] was proposed mainly for detection of traffic signs but was found partially successful by the same authors for recognition too. An ant colony optimization based method [49] performed well for natural category of occlusions for both datasets but produced poor results on other types of occlusions. Since this method employs an ant colony optimization based feature selection technique, it can be thought of as the closest to our proposed techniques. However, this method was designed for standard traffic signs whereas our proposed work targets traffic signs partially occluded by either other nearby objects or suffered with degradations.

Table 3: Comparing “Recognition Accuracy” of various methods on both datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset1</td>
<td>Natural</td>
<td>0.71</td>
<td>0.68</td>
<td>0.60</td>
<td>0.71</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Incidental</td>
<td>0.64</td>
<td>0.58</td>
<td>0.63</td>
<td>0.65</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Deliberate</td>
<td>0.49</td>
<td>0.55</td>
<td>0.45</td>
<td>0.56</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Degraded</td>
<td>0.70</td>
<td>0.69</td>
<td>0.54</td>
<td>0.66</td>
<td>0.81</td>
</tr>
<tr>
<td>Dataset2</td>
<td>Natural</td>
<td>0.75</td>
<td>0.77</td>
<td>0.65</td>
<td>0.75</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>Incidental</td>
<td>0.65</td>
<td>0.45</td>
<td>0.55</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Deliberate</td>
<td>0.48</td>
<td>0.52</td>
<td>0.51</td>
<td>0.65</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Degraded</td>
<td>0.67</td>
<td>0.55</td>
<td>0.61</td>
<td>0.46</td>
<td>0.86</td>
</tr>
</tbody>
</table>

5.2. Effect of Dimensionality Reduction

In order to know whether feature interaction based dimensionality reduction technique proposed in Section 3.5 is effective to enhance recognition accuracy and to reduce processing time, the whole process described in Section 3 was repeated with and without dimensionality reduction. The results given in Table 4 show that by employing feature interaction based dimensionality reduction, we achieved a better average recognition accuracy on occluded and degraded traffic signs found in both datasets. Further, working on data matrix of reduced features i.e., $S$, saves computational time too. It is shown in Table 4 that with a reduced subset of only about 10% of features, we are able to obtain the accuracy achieved with 100% features. Moreover, more than 80% saving in execution time was also achieved for both datasets.

Table 4: Effect of using reduced feature set as a result of dimensionality reduction technique used in Section 3.5 on recognition accuracy and processing time.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Recognition Accuracy (No. of Features)</th>
<th>Time on $S$(on $D$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full Feature Set</td>
<td>Reduced Feature Set</td>
</tr>
<tr>
<td>Dataset1</td>
<td>0.79 (5,376)</td>
<td>0.81 (1,885)</td>
</tr>
<tr>
<td>Dataset2</td>
<td>0.76 (5,376)</td>
<td>0.80 (1,915)</td>
</tr>
</tbody>
</table>

5.3. Amount of Occlusion versus Average Error Rate

It is obvious that increasing amount of occlusions cause error rate to increase. We performed experiments such that degree of occlusion applied on a traffic signs was increased from 5% to 50% (by synthetically putting real world occlusions on standard traffic signs) and the error rate was computed for every sample. The results show an increasing trend in error rate for individual signs belonging to each category i.e., Natural, Incidental and Deliberate. Finally, an average error rate was computed for each category of signs found in both datasets and is shown in Figure 13. It is obvious that increasing percentage of occlusion increases error rate for all four categories of occlusions and for both datasets.
Figure 13: Average error rate of images from natural, incidental and deliberate categories for dataset1 (a) and dataset2 (b).
5.4. Computational Complexity

The proposed technique was also compared with other state of the art and recent traffic sign recognition methods in terms of the computational complexity. For this comparison, we have selected four methods, three of which partially address the problem of partial occlusion in traffic sign recognition [26, 35, 27] and the fourth incorporates an ant colony optimization based feature selection technique [49]. The comparison given in Table 5 shows that all four methods have comparable computational complexity. However, the proposed method works just better than all other techniques in addition to being the most accurate as demonstrated in the results given in Tables 2 and 3.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset 1</td>
<td>1.2</td>
<td>1.5</td>
<td>1.4</td>
<td>.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Dataset 2</td>
<td>1.1</td>
<td>1.3</td>
<td>1.5</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

6. Discussion of Results

It was demonstrated in Section 5 that our proposed method outperformed various other techniques on both datasets by achieving maximum precision, recall and recognition accuracy values. Mainly this advantage in performance is due to (1) the computation of a boosted image in which the effect of occlusion is suppressed (2) our proposed recognition method takes maximum benefit by continuously updating the underlying statistical model even during testing and (3) the use of dimensionality reduction increases both computational complexity and the overall accuracy.

An advantage achieved by our proposed method in precision and recall values (as shown in Table 2) in comparison of various other techniques is mainly due to the computation of boosted image and the use of our novel CNN inspired DCT phase based multiscale feature extraction technique. On the other hand, the other techniques utilize either grayscale or selected channel of the RGB input image. The highest recognition accuracy for first three types of occlusions mentioned in Table 3 is by virtue of our proposed feature extraction technique. However, for the fourth type i.e. degraded, this is due to adaptive computation of the boosted image.

Further, the dimensionality reduction algorithm proposed in this work involves computing mutual information and feature interaction. Both these operations involve joint probability terms and hence are computationally expensive. Since our descriptor contains only the binary values (either +1 or -1), computing joint probability terms becomes easier. Further, these computations are to be performed only during training and the resulting advantage is reduced execution time for test set.

It is shown in Figure 13 that the error rate for all categories of occlusions increases versus increasing amount of occlusion but the Natural category is least affected. This is because for this category, both color and texture of pixels belonging to occlusions are different from that of traffic signs i.e., most of the plants are green and the repeating patterns (textures) on these occlusions have non uniform distribution of pixels. It is well known that for traffic signs, colors are bright and their distribution throughout the traffic sign is uniform. This characteristic of natural category of occlusion makes it least prone to error. Occlusions of type Incidental and Deliberate are more likely to be affected by increasing amount of occlusions. Not surprisingly, this behavior remains the same for both datasets.

Finally, a number of correctly and incorrectly recognized signs by our proposed method are given in Figure 14. It can be seen that the proposed method fails as the occlusions cover more than 50% or the contents are completely suppressed.
7. Conclusion and Future Work

In this paper, we have proposed a strategy to recognize occluded traffic signs. The task is challenging because a part of the traffic sign is suppressed by another object or the sign is weirded or degraded. We identified different categories of occlusions found in real world cases namely natural, incidental, deliberate and degradations. Given a detected partially occluded traffic sign image, we first compute a boosted image that by, taking inspiration from human cognition; suppressing any occluded or degraded portions and highlighting only the visible areas pertaining to the traffic sign. The statistical model used to compute boosted image is dynamic i.e., the parameters are continuously updated as the system is tested in the field. To extract invariant features, we propose a convolutional neural network inspired multiscale spectral method utilizing phase spectrum of discrete cosine transform. To remove irrelevant and redundant information, an interaction based dimensionality reduction method was employed to improve recognition accuracy and save execution time. Experiments were conducted on two datasets namely Dataset1 and Dataset2. Dataset1 is composed of naturally occluded images collected from authors’ country of residence in addition to synthetically occluded samples. Dataset2 was derived from famous German Traffic Sign Recognition Benchmark (GTSRB) and contains originally as well as synthetically generated images. Experimental results on both datasets show that our proposed method outperforms state of the art methods on both datasets. Moreover, the proposed dimensionality reduction technique showed outstanding savings in time and demonstrates significant increase in accuracy.

As a future work, we are planning to investigate the performance of convolutional neural networks (CNN) on datasets comprising of occluded and degraded traffic signs. Up till now CNN based methods reported in literature aim detection and recognition of standard traffic signs only. Since training a deep CNN is computationally expensive and is prone to over fitting for small datasets comprising only of a few thousand images, use of transfer learning to handle occlusions can be a
possible future work. Another future effort is to test the adaptive parameter estimation in real world scenarios for a longer duration of time and analyzing the behavior in case the amount of field/test data exceeds the initial training data.

References


[34] Ishida, H., Takahashi, T., Ide, I., Mekada, Y. and Murase, H. “Identification of degraded traffic sign symbols by a generative learning method,” *18th International Conference on Pattern Recognition (ICPR)*, Hong Kong, China, pp. 531–534 (2006).


Abdul Mannan received his BSc, MSc and PhD degrees in Electrical Engineering from University of Engineering & Technology, Lahore, Pakistan in 2002, 2007 and 2021 respectively. Right now, he is working as Assistant Professor at Department of Electrical Engineering at NFC Institute of Engineering & Technology, Multan, Pakistan. His research interests include machine learning, pattern recognition, image processing and computer vision.

Kashif Javed received the BSc, MSc, and PhD degrees in Electrical Engineering in 1999, 2004, and 2012, respectively, from the University of Engineering and Technology (UET), Lahore, Pakistan. He joined the Department of Electrical Engineering at UET in 1999 where he is currently a professor. He has been a reviewer of IEEE Transactions on Knowledge and Data Engineering (TKDE) since 2011. His research interests include machine learning, pattern recognition, and natural language processing for Urdu.

Atta ur Rehman received BS degree in electrical engineering from Air university, Islamabad, Pakistan and PhD degree in electrical engineering from Loughborough university, UK. Currently, he is with National University Science and Technology (NUST), Islamabad, Pakistan. His research interests are image processing, machine learning and pattern recognition.

Haroon A. Babri received the BSc degree in Electrical Engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, in 1981, and the MS and PhD degrees in Electrical Engineering from the University of Pennsylvania in 1991 and 1992, respectively. He was with the Nanyang Technological University, Singapore, from 1992 to 1998, with the Kuwait University from 1998 to 2000, and with the Lahore University of Management Sciences (LUMS) from 2000 to 2004. He is currently a professor of electrical engineering at UET. He has written two book chapters and has more than 60 publications in machine learning, pattern recognition, neural networks, and software reverse engineering.

Serosh K. Noon received her MSc degree in Electrical Engineering from University of Engineering and Technology, Lahore, Pakistan, and is presently working towards her PhD degree in Electrical Engineering. She is also with the Department of Electrical Engineering at NFC Institute of Engineering and Technology, Multan, Pakistan. Her research interests include biomedical image and signal processing, artificial intelligence in agriculture, and computer vision.