International Arab Conference on Information Technology (ACIT'2016)

Towards Accurate Real-Time Traffic Sign Recognition Based on Unsupervised Deep Learning of Spatial Sparse Features: A Perspective

Ahmad M. Hasasneh[1], Yousef-Awwad Daraghmi[2], and Nabil M. Hasasneh[3]
[1]Department of Information Technology, Palestine Ahliya University, Palestine
[2]Computer Engineering Department, Palestine Technical University, Palestine
[3]Computer Science Department, Hebron University, Palestine

Abstract: *Learning a good generative model is of utmost importance for the problems of computer vision, image classification and image processing. In particular, learning features from small tiny patches and perform further tasks, like traffic sign recognition, can be very useful. In this paper we propose to use Deep Belief Networks, based on Restricted Boltzmann Machines and a direct use of tiny images, to produce an efficient local sparse representation of the initial data in the feature space. Such a representation is assumed to be linearly separable and therefore a simple classifier, like softmax regression, is suitable to achieve accurate and fast real-time traffic sign recognition. However, to achieve localized features, data whitening or at least local normalization is a prerequisite for these approaches. The low computational cost and the accuracy of the model enable us to use the model on smart phones for accurately recognizing traffic signs and alerting drivers in real time. To our knowledge, this is the first attempt that tiny images feature extraction using deep architecture is a simpler alternative approach for traffic sign recognition that deserves to be considered and investigated.*

Keywords: *Traffic Sign Recognition, Image Processing, Image Classification, Computer Vision, Restricted Boltzmann Machines, Deep Belief Networks, Softmax Regression, Sparse Representation.*

1. Introduction

As computers offer more and more processing power, the aim of real time traffic sign detection and recognition is becoming visible. Traffic signs detection and recognition provide drivers important information that makes driving safe and convenient. Some models of high class vehicles already come equipped with driver assistance systems which provide automated detection and recognition of certain classes of traffic signs. In driver assistance systems or road inventory systems, the problem is no longer how to efficiently detect and recognize a traffic sign in a single image but how reliable and efficient the detection method detect the huge video frames without any false alarms.

Detecting and recognition an object in an image is a computer vision problem for which a wide variety of algorithms exist. For example, Edge histogram information is used in traffic sign detection [1], however the mechanism used in this paper is specially designed for certain shapes signs and cannot be extended to more general signs easily. In addition, Timofte et al. [32] present a complex system for traffic sign detection and recognition. In this proposed work an efficient algorithm and methods based on new machine learning can be used to achieve automatic alert traffic signs detection and recognition. Several studies have proposed traffic sign recognition (TSR) systems based on different image processing and machine learning algorithms. However, the efficiency of the proposed TSR algorithms requires improvement to enable real-time alerts on onboard devices which have limited computational power. Further improvement on accuracy of TSR is also required mainly in unstable weather conditions or when multiple signs exist on one pillar.

The rest of the paper is organized as follows. Section 2 presents the current approaches to TSR. Section 3 describes the proposed model to achieve TSR. Conclusions and future works are presented in section 4.

2. Current Approaches

Several related and significant studies stated that the electronic identification of road signs requires two main phases: detection and recognition [19, 22, 30, 31, 37]. In the detection phase, a traffic sign can be detected either by shape or color. Shape based detection algorithms use techniques such as the Hough transform to detect lines and identify the shape of the road sign [4, 28]. However, the existence of non-traffic similar shapes on roads affects the accuracy of this algorithm. Further, the extraction of shape usually consumes large computational time. On the other hand, color based algorithms are a color segmentation

of Red, Green and Blue regions in the given image [7, 8, 34, 38]. These algorithms are good enough for segmenting traffic signs in ideal illumination condition. Further these algorithms need defining many threshold values for the colors. Color based detection employs transforming images to HIS or HSV color spaces. The transformation is computationally demanding and therefore researchers moved to the RGB color space in order to speed up the detection procedure.

Regarding to classification of colors or shapes, the Support Vector Machine (SVM) is often used as in [10, 23]. The SVM shows good accuracy when images are rotated. Other classification approaches include the Artificial Neural Networks (ANN) as in [34]. The normalized correlation-based pattern matching was also used to classify signs based on colors [11]. In these approaches, it is possible to represent regions of interest based on illumination values (pixel based approaches) or on image features (feature based approach). Despite the good accuracy of the TSR that utilize these classification techniques, their computation performance is not suitable for real-time recognition on onboard devices which have limited computational resources.

The K-nearest neighborhood (KNN) has been also used in traffic sign classification [21]. The KNN is a machine learning technique, and it is widely used in pattern classification. In some cases, the KNN outperforms the SVM for real time TSR system. The KNN is borrowed to our research for increasing the accuracy and reducing the computational time.

After presenting the different existing approaches that have been used to achieve TSR, we have noted that these approaches generally include two main phases of coding and classification. We have also seen that most of the coding methods are based on hand-crafted feature extractors, which are empirical detectors. By contrast, a set of recent methods based on deep architectures of neural networks give the ability to build it from theoretical considerations.

TSR therefore requires projecting images onto an appropriate feature space that allows an accurate and rapid classification. Contrarily to these empirical methods mentioned above, new machine learning methods have recently emerged which strongly related to the way natural systems code images [25]. These methods are based on the consideration that natural image statistics are not Gaussian as it would be if they have had a completely random structure [9]. The auto-similar structure of natural images allowed the evolution to build optimal codes. These codes are made of statistically independent features and many different methods have been proposed to construct them from image datasets. Imposing locality and sparsity constraints in these features is very important. This is probably due to the fact that any simple algorithms based on such constraints can achieve linear signatures similar to the notion of receptive field in natural systems. Recent years have seen an interesting interest in computer vision algorithms that rely on local sparse image representations, especially for the problems of image classification and object recognition [5, 12, 26, 35, 36]. Moreover, from a generative point of view, he effectiveness of local sparse coding, for instance for image reconstruction [18], is justified by the fact that a natural image can be reconstructed by a smallest possible number of features. It has been shown that Independent Component Analysis (ICA) produces localized features. Besides it is efficient for distributions with high kurtosis well representative of natural image statistics dominated by rare events like contours; however the method is linear and not recursive. These two limitations are released by Deep Belief Networks (DBNs) [15] that introduce non-linearities in the coding scheme and exhibit multiple layers. Each layer is made of a Restricted Boltzmann Machine (RBM), a simplified version of a Boltzmann machine proposed by Smolensky [29] and Hinton [14]. Each RBM is able to build a generative statistical model of its inputs using a relatively fast learning algorithm, Contrastive Divergence (CD), first introduced by Hinton [14]. Another important characteristic of the codes used in natural systems, the sparsity of the representation [25], is also achieved in DBNs. Moreover, it has been shown that these approaches remain robustness to extract local sparse efficient features from tiny images [33]. This model has been successfully used in [12] to achieve semantic place recognition. The hope is to demonstrate that DBNs coupled with tiny images can also be successfully used in the context of TSR.

3. Proposed Model

The methodology of this research mainly includes four stages (see figure 1) which can be summarized as follows: 1) data collection and image acquisition, 2) image pre-processing and whitening, 3) feature extraction and image coding and finally 4) traffic sign recognition.

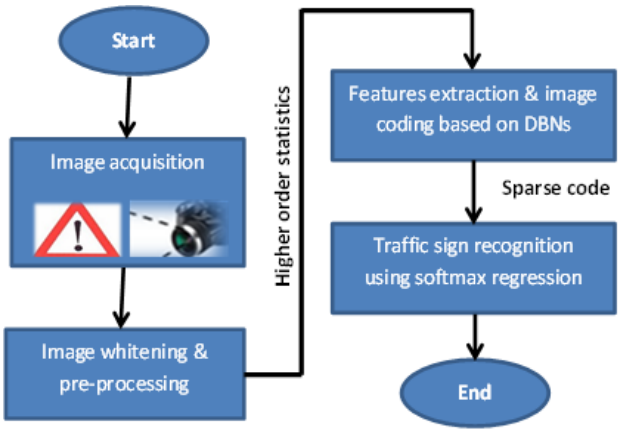

Figure 1. Proposed model stages.

3.1 Image Pre-processing

The typical input dimension for a DBN is approximately 1000 units (e.g. 30x30 pixels). Dealing with smaller patches could make the model unable to extract interesting features. Using larger patches can be extremely time-consuming during feature learning. Additionally the multiplication of the connexion weights acts negatively on the convergence of the CD algorithm. The question is therefore how could we scale the size of realistic images (e.g. 300x300 pixels as shown in figure 2) to make them appropriate for DBNs?

Figure 2. Left: Examples of some traffic signs of different colors. Right: The corresponding tiny images of the traffic signs. One can see that, despite the size reduction, these small images remain fully recognizable.

Three solutions can be envisioned. The first one is to select random patches from each image as done in [27], the second is the use of convolutional architectures, as proposed in [20], and the last one is to reduce the size of each image to a tiny image as proposed in [33]. The first solution extracts local features and the characterization of an image using these features can only be made using BoWs approaches we wanted to avoid. The second solution shows the same limitations as the first one and additionally gives raise to extensive computations that are only tractable on Graphics Processing Unit architectures. Features extraction using random patches is irrespective of the spatial structures of each image [24]. In the case of structured scenes like the ones used in TSR these structures bear interesting information.

Besides, tiny images have been successfully used in [33] for classifying and retrieving images from the 80-million images database developed at MIT. Torralba in [33] showed that the use of tiny images combined with a DBN approach led to code each image by a small binary vector defining the elements of a feature alphabet that can be used to optimally define the considered image. The binary vector acts as a bar-code while the alphabet of features is computed only once from a representative set of images. The power of this approach is well illustrated by the fact that a relatively small binary vector largely exceeds the number of images that have to be coded even in a huge database ($2^{256} \approx 10^{75}$). So, for all these reasons we have chosen image reduction.

On the other hand, natural images are highly structured and contain significant statistical redundancies, e.g. their pixels have strong correlations [2, 3]. Removing these correlations is known as whitening. It has been shown that whitening is a mandatory step for the use of clustering methods in object recognition [6]. Whitening being a linear process and it does not remove the higher order statistics present in the data.

As a consequence, as proposed by [12, 33], after color conversion and image cropping, the image size is reduced to 42x24 as shown in figure 2, right. The final set of tiny images will be centred and whitened in order to eliminate order 2 statistics. Consequently the variance in equation 6 will be set to 1. Contrarily to [33], the 42x24 = 1008 pixels of the whitened images will be used directly as the input vector of the network for features extraction purpose.

3.2 Gaussian-Bernoulli Restricted Boltzmann Machines

Unlike a classical Boltzmann Machine, a RBM is a bipartite undirected graphical model $\theta = \{w_{ij}, b_i, c_j\}$, linking, through a set of weights w_{ij} between visible and hidden units and biases $\{b_i, c_j\}$ a set of visible units **v** to a set of hidden units **h** [6]. For a standard RBM, a joint configuration of the binary visible units and the binary hidden units has an energy function, $E(b_i, c_j, \theta)$ given by:

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i,j} v_i h_j w_{ij} - \sum_i b_i v_i - \sum_j c_j h_j \quad (1)$$

The probabilities of the state for a unit in one layer conditional to the state of the other layer can therefore be easily computed. According to Gibbs distribution:

$$P(\mathbf{v}, \mathbf{h}; \theta) = -\frac{1}{Z(\theta)} e^{-E(\mathbf{v},\mathbf{h};\theta)} \quad (2)$$

where $Z(\theta)$ is a normalizing constant. Thus after marginalization:

$$P(\mathbf{h};\theta) = \sum_{\mathbf{v}} P(\mathbf{v},\mathbf{h};\theta) \quad (3)$$

it can be derived [17] that the conditional probabilities of a standard RBM are given as follows:

$$P(\mathrm{h_j} = 1|\boldsymbol{v};\theta) = \sigma\left(\mathrm{c_j} + \sum_i w_{ij}\, v_i\right) \quad (4)$$

$$P(\mathrm{v_i} = 1|\boldsymbol{h};\theta) = \sigma\left(\mathrm{b_i} + \sum_j w_{ij}\, \mathrm{h_j}\right) \quad (5)$$

where $\sigma(x) = 1/(1+e^{-x})$is the logistic function.

However, logistic or binary visible units are not appropriate for multi-valued inputs like pixel levels, because logistic units are a very poor representation for data such as patches of natural images. To overcome this problem, as suggested by [16], in the present work we will replace the binary visible units by a zero-mean Gaussian activation scheme as follows:

$$P(\mathrm{v_i} = 1|\boldsymbol{h};\theta) \leftarrow \mathrm{N}\left(\mathrm{b_i} + \sum_j w_{ij}\, \mathrm{h_j}, \sigma^2\right) \quad (6)$$

In this case the energy function of Gaussian-Bernoulli RBM is given by:

$$E(\mathbf{v},\mathbf{h};\theta) = \sum_i \frac{(v_i - \mathrm{b_i})^2}{2\sigma_i^2} - \sum_j c_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} \quad (7)$$

3.3 Learning RBM Parameters

One way to learn RBM parameters is through the maximization of the model log-likelihood in a gradient ascent procedure. The partial derivative of the log-likelihood for an energy-based model can be expressed as follows:

$$\frac{\partial}{\partial\theta} L(\theta) = -\langle \frac{\partial E(\boldsymbol{v},\theta)}{\partial\theta} \rangle_{data} + \langle \frac{\partial E(\boldsymbol{v},\theta)}{\partial\theta} \rangle_{model} \quad (8)$$

where $\langle\,\rangle_{model}$ is an average with respect to the model distribution and $\langle\,\rangle_{data}$ an average over the sample data. The energy function of a RBM is given by:

$$E(\boldsymbol{v},\theta) = log \sum_{\boldsymbol{h}} exp^{-E(\boldsymbol{v},\boldsymbol{h},\theta)} \quad (9)$$

and

$$\partial E(\boldsymbol{v},\theta)/\partial\theta = \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v};\theta)\partial E(\boldsymbol{v},\boldsymbol{h},\theta)/\partial\theta \quad (10)$$

Unfortunately, computing the likelihood needs to compute the partition function, Z(θ), that is usually intractable. However, Hinton [17] proposed an alternative learning technique called CD. This learning algorithm is based on the consideration that minimizing the energy of the network is equivalent to minimize the distance between the data and a statistical generative model of it. A comparison is made between the statistics of the data and the statistics of its representation generated by Gibbs sampling. Hinton [17] showed that usually only a few steps of Gibbs sampling (most of the time reduced to one) are sufficient to ensure convergence. For a RBM, the weights of the network can be updated as follows:

$$w_{ij} \leftarrow w_{ij} + \eta\left(\langle v_i^0 h_j^0\rangle - \langle v_i^n h_j^n\rangle\right) \quad (9)$$

Where η is the learning rate, v^0 corresponds to the initial data distribution, $h^0$is computed using equation 4, v^n is sampled using the Gaussian distribution in equation 6 and with n full steps of Gibbs sampling, and h^n is again computed from equation 4.

3.4 Layerwise Training for Deep Belief Networks

A DBN is a stack of RBMs trained in a greedy layer-wise and bottom-up fashion introduced by [15]. The first model parameters are learned by training the first RBM layer using the contrastive divergence. Then, the model parameters are frozen and the conditional probabilities of the first hidden unit values are used to generate the data to train the higher RBM layers. The process is repeated across the layers to obtain a sparse representation of the initial data that will be used as the final output.

3.5 Traffic Sign Recognition

Assuming that the non-linear transform operated by DBN improves the linear separability of the data, a simple regression method will be used to achieve TSR. To express the final result as a probability that a given sign means one thing, we normalize the output with a softmax regression method. According to maximum likelihood principles, the largest probability value gives the decision of the system.

The classification process will also be investigated using a more sophisticated classifier, a SVM classification method instead of softmax regression. In case of comparable results; this will underline that the DBN computes a linear separable signature of the initial data and it can be an alternative approach to achieve TSR that can be deserved to be considered and investigated.

4. Conclusions and Future Works

The aim of this paper is therefore to propose to use DBNs coupled with tiny images in a challenging image recognition task, view-based TSR. The

expected results should demonstrate that an approach based on tiny images followed by a projection onto an appropriate feature space can achieve interesting classification results in an TSR task. Our hope is get comparable results or even to outperform the results obtained in [10, 21] based on more complex techniques. In case of comparable results, this paper is thus offer a simpler alternative to the method recently proposed in [10, 21] based on cue integration and the computation of a confidence criterion in a SVM or a KNN classification approach.

Our future work is to empirically investigate the proposed model to achieve traffic sign recognition. The first step is to perform the whitening process and study its influences on the final classification results. Learning a set of features and use them to create an appropriate code of the initial data that could simplify the overall classification process. Finally, assuming that the non-linear transform operated by DBN improves the linear separability of the data, a simple regression method will be used to perform the classification process. The classification process will also be examined using sophisticated classification techniques like SVM in order to investigate whether the linear separability is gained by DBN or not.

The proposed system will first classify the traffic signs into four categories, namely: triangles representing the warning signs, rectangles, which represent the information signs, circles which represent the regulatory signs and finally traffic light signs. This research can therefore be extended to achieve TSR based on DBNs and Contrastive Divergence which has been used to achieve semantic place recognition [13]. To our knowledge, this approach has not been investigated yet which deserves to be considered.

Moreover, our future work will also include prioritizing traffic signs particularly when more than one sign are detected. Prioritizing can be performed based on the distance between the car and the sign or based on the importance of the sign. This would make our system more intelligent by giving alerts to drivers according to the importance of the sign.

References

[1] Alefs B., Eschemann G., Ramoser H., and Beleznai C., "Road sign detection from edge orientation histograms," *in Intelligent Vehicles Symposium*, 2007 IEEE. IEEE, 2007, pp. 993–998.

[2] Attneave F., "Some informational aspects of visual perception." *Psychological review*, vol. 61, no. 3, p. 183, 1954.

[3] Barlow H., "Redundancy reduction revisited," *Network: computation in neural systems*, vol. 12, no. 3, pp. 241–253, 2001.

[4] Belaroussi R. and Tarel J.-P., "Angle vertex and bisector geometric model for triangular road sign detection," *in Applications of Computer Vision (WACV)*, 2009 Workshop on. IEEE, 2009, pp. 1–7.

[5] Boureau Y.-L., Bach F., LeCun Y., and Ponce J., "Learning mid-level features for recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2559–2566.

[6] Coates A., Ng A. Y., and Lee H., "An analysis of single-layer networks in unsupervised feature learning," *in International conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[7] De la Escalera A., Armingol J. M., and Mata M., "Traffic sign recognition and analysis for intelligent vehicles," *Image and vision computing*, vol. 21, no. 3, pp. 247–258, 2003.

[8] De La Escalera A., Moreno L. E., Salichs M. A., and Armingol J. M., "Road traffic sign detection and classification," *IEEE Transactions on Industrial Electronics*, vol. 44, no. 6, pp. 848–859, 1997.

[9] Field D. J., "What is the goal of sensory coding?" *Neural computation*, vol. 6, no. 4, pp. 559–601, 1994.

[10] Gomez-Moreno H., Maldonado-Bascon S., Gil-Jimenez P., and Lafuente-Arroyo S., "Goal evaluation of segmentation algorithms for traffic sign recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 4, pp. 917–930, 2010.

[11] Greenhalgh J. and Mirmehdi M., "Real-time detection and recognition of road traffic signs," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1498–1506, 2012.

[12] Hasasneh A., Frenoux E., and Tarroux P., "Semantic place recognition based on deep belief networks and tiny images." *in ICINCO (2)*. SciTePress, 2012, pp. 236–241.

[13] Hasasneh A., *Robot semantic place recognition based on deep belief networks and a direct use of tiny images*, Ph.D. dissertation, University of Paris SUD XI, Paris, 2012.

[14] Hinton G. E., "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[15] Hinton G. E., S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] Hinton G., "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[17] Krizhevsky A., "Learning multiple layers of features from tiny images," *Tech. Rep.*, 2009.

[18] Labusch K. and Martinetz T., "Learning sparse odes for image reconstruction." *in ESANN*, 2010.

[19] Lai C.-H. and Yu C.-C., "An efficient real-time traffic sign recognition system for intelligent vehicles with smart phones," *International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, IEEE, 2010, pp. 195–202.

[20] Lee H., Grosse R., Ranganath R., and Ng A. Y., "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," *in Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[21] Lillo-Castellano J., Mora-Jimenez I., Figuera-Pozuelo C., and Rojo-Alvarez J., "Traffic sign segmentation and classification using statistical learning methods," *Neurocomputing*, vol. 153, pp. 286–299, 2015.

[22] Liu H., Liu Y., and Sun F., "Traffic sign recognition using group sparse coding," *Information Sciences*, vol. 266, pp. 75–89, 2014.

[23] Miura J., Kanda T., and Shirai Y., "An active vision system for real-time traffic sign recognition," *in Intelligent Transportation Systems*, 2000. Proceedings. 2000 IEEE. IEEE, 2000, pp. 52–57.

[24] Norouzi M., Ranjbar M., and Mori G., "Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning," *in Computer Vision and Pattern Recognition*, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 2735–2742.

[25] Olshausen B. A. and Field D. J., "Sparse coding of sensory inputs," *Current opinion in neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[26] Ranzato M. A., Huang F. J., Boureau Y.-L., and LeCun Y., "Unsupervised learning of invariant feature hierarchies with applications to object recognition," *in Computer Vision and Pattern Recognition, CVPR'07*. IEEE Conference on. IEEE, 2007, pp. 1–8.

[27] Ranzato M. A., Krizhevsky A., and Hinton G. E., "Factored 3-way restricted Boltzmann machines for modeling natural images," *in International Conference on Artificial Intelligence and Statistics*, 2010, pp. 621–628.

[28] Sallah S. S. M., Hussin F. A., and Yusoff M. Z., "Shape-based road sign detection and recognition for embedded application using Matlab," *in Intelligent and Advanced Systems (ICIAS)*, 2010 International Conference on. IEEE, 2010, pp. 1–5.

[29] Smolensky P., "Information processing in dynamical systems: Foundations of harmony theory," *in Parallel Distributed Processing: Volume 1: Foundations*, D. E. Rumelhart, J. L. McClelland et al., Eds. Cambridge: MIT Press, 1987, pp. 194–281.

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

[31] Sun Z.-L., Wang H., Lau W.-S., Seet G., and Wang D., "Application of bw-elm model on traffic sign recognition," *Neurocomputing*, vol. 128, pp. 153–159, 2014.

[32] Timofte R., Zimmermann K., and Van Gool L., "Multi-view traffic sign detection, recognition, and 3d localisation," *Machine Vision and Applications*, vol. 25, no. 3, pp. 633–647, 2014.

[33] Torralba A., Fergus R., and Weiss Y., "Small codes and large image databases for recognition," *IEEE Conference on in Computer Vision and Pattern Recognition (CVPR 2008)*. IEEE, 2008, pp. 1–8.

[34] Vitabile S., Pollaccia G., Pilato G., and Sorbello F., "Road signs recognition using a dynamic pixel aggregation technique in the hsv color space," *11th IEEE International Conference on Image Analysis and Processing*, 2001. Proceedings, 2001, pp. 572–577.

[35] Wright J., Ma Y., Mairal J., Sapiro G., Huang T. S., and Yan S., "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, 2010.

[36] Yang J., Yu K., Gong Y., and Huang T., "Linear spatial pyramid matching using sparse coding for image classification," *in Computer Vision and Pattern Recognition*, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 1794–1801.

[37] Zaklouta F. and Stanciulescu B., "Real-time traffic sign recognition in three stages," *Robotics and autonomous systems*, vol. 62, no. 1, pp. 16–24, 2014.

[38] Zhu S. and Liu L., "Traffic sign recognition based on color standardization," *IEEE International Conference on Information Acquisition*, 2006, pp. 951–955.