Traffic Sign Recognition using Generative Model of Scale Invariant Feature Descriptors

Sai Sankalp Arrabolu
Department of Electrical Engineering
Indian Institute of Technology
Roorkee - 247667 Uttarakhand, India
cyberuee@iitr.ernet.in

Lucas Paletta
Institute of Digital Image Processing
Joanneum Research Forschungsgesellschaft mbH
Graz, A-8010, Austria
lucas.paletta@joanneum.at

***Abstract*— This paper presents an approach based on a generative model of Scale Invariant Feature Transform (SIFT) descriptors for the purpose of traffic sign recognition. In our work, SIFT key descriptors [1] are grouped and therefore providing rather unique descriptors of traffic signs by referencing the descriptor's main orientation to the center of the sign and thus the single SIFT based keypoint matching is significantly improved. A careful analysis of the performance of the descriptor grouping on the real world traffic sign imagery is presented. Measures have been taken for preventing ambiguity within the voting for different sign hypotheses. Since the approach uses SIFT features it is rather invariant to image scaling, translation, illumination changes and affine projections. The approach shows good performance on a wide range of images with different scales, in plane rotations and partial occlusions. This approach allows reliable detection of multiple traffic signs of different categories in the same image. The database used for the training of the nearest neighbor classifier consists of real world traffic signs. The performance in the detection of traffic signs of images from the IMSERV database proved to be around 80% in accuracy. The performance of this robust approach is competitive to the state-of-art approaches that are dedicated to the traffic sign recognition problem.**

***Index Terms*—Nearest Neighbor Classifier, SIFT Key Descriptors.**

I. INTRODUCTION

The Traffic Sign Recognition has been a concern in computer vision research for the past few years. The aim is to create an automatic detection and classification of traffic signs from the real world traffic scene images for driving assistance systems (DAS) and mobile mapping applications. The traffic signs describe the current traffic scenario, define the right of way, prohibit or permit certain directions, warn about risky factors etc. It is desirable to design a smart car control system to prevent the hazards of careless driving and automatic traffic sign detection will provide a huge step towards it. There have been many approaches for the traffic sign recognition. Some of them are based on color images processing [2], color thresholding techniques [3], shape features [4] [5], distance transforms [6]. The problem with the color based approach is that the colors of the road sign may fade away after long exposure to the sun, the paint may peel away, and that color appearance dramatically changes with various illuminations. There are various neural network based solutions that confront this problem [7] [3] [8].

Some of the main problems faced in traffic sign recognition are dependence of recognition on the scale of the image, the rotation, poor lighting conditions (lighting is different according to the time of the day, season, cloudiness and other weather conditions), partial occlusions, shadows, perspective distortions, etc. For dealing with these problems we pool the main ideas reported in the literature [1] [9] [10] [11]. We use a lucid approach rather than adopting one of the discriminative methods based on machine learning techniques. In our approach the SIFT descriptors of a particular trained traffic sign are made unique by calculating the distance of that descriptor from the center of the image and the angle between the line joining the center of the image to the descriptor and its original orientation (Sec. II). The search for the nearest neighbors is done using a kd-tree approach. During the testing, for all the descriptors of the test image the n nearest neighbors are obtained. Each of the test descriptor is relocated by using the distance and the angle data of every nearest neighbor obtained from the kd-tree. As a result, SIFT descriptors voting for a particular sign within the trained database are clustered at the center of the traffic sign in the test image. This clustering enables us to detect the traffic signs present in the test image. Since we use the Scale Invariant Feature approach, the database (Sec. III) used for the training consists of real world images.

II. METHODOLGY

In this Section we present the theory and formal mathematical background used in our approach. We first present the feature extraction, the calculation of the distance and orientation of the descriptor with respect to the center of the image and the search algorithm for the nearest neighbors. Next we explain about the procedure followed for the

detection and the recognition of the traffic signs.

Image Features. Our approach is based on Scale Invariant Feature Transform (SIFT) proposed in [1]. This method provides scale, position, dominant orientation of a feature with respect to its neighborhood and a 128 dimensional descriptor based on local gradient information computed with respect to the dominant orientation of the feature. In addition to these parameters we add two additional parameters (d, Ø) for every feature that refer to the orientation Ø and the distance d of the descriptor with respect to the centre of the traffic sign. Since the shape of a traffic sign is the most prominent feature, we have reduced the threshold on the minimum contrast in order to obtain more features on the boundary of the traffic signs

Orientation and Distance with respect to Centre. In [12] [13] [14] the position of the features is related to the reference frame or object center in Cartesian coordinates. This restricts it to one pose of the object. As stated in [10] we express the position of the SIFT feature with respect to the center in polar coordinates. Fig. 1 clearly illustrates how the orientation and the distance are calculated. The mathematical relations for the calculation of distance and angle are as follows

$$d = \sqrt{(x_{sift} - x_{centre})^2 + (y_{sift} - y_{centre})^2} \qquad (1)$$

$$\alpha = \arctan\left(y_{sift} - y_{centre}\right) / \left(x_{sift} - x_{centre}\right)$$

$$\phi = \begin{cases} \pi - \alpha - \theta & (if\ x_{sift} > x_{centre}) \\ -\alpha - \theta & (otherwise) \end{cases} \qquad (2)$$

The angles Ø and θ are shown in Fig.1.

Kd-tree. The SIFT features are found for all the trained images present in the database. Every feature belonging to a specific traffic sign is assigned a label corresponding to that particular sign. The features along with the labels are given as input to a kd-tree categorizer for supervised learning. This improves the search efficiency for finding the nearest neighbors of the query descriptor.

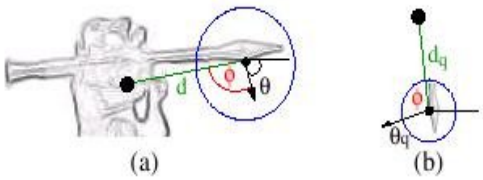

Figure 1: (a) The position of the descriptor with respect to the center, d and Ø represent the position and orientation with respect to the centre. (b) Retracing the center from the descriptor of the test image. The distance d_q is drawn depending on the scale of the image.

Every trained SIFT descriptor is placed at the node of the kd-tree (Fig. 2). The kd-tree used in the approach is an ANN (Approximate Nearest Neighbor) data structure which is based on a recursive subdivision of space into disjoint hyper rectangular regions called cells. Each node of the tree is associated with a box, and is associated with a set of data points, i.e. the SIFT descriptors that lie within this box.

The standard recursive search of ANN is adopted for the nearest neighbor search. When the first node of the tree is encountered the algorithm visits the first descriptor which is closest to the query point. If the box containing the other leaf lies within 1/1+ Є, (Є being a positive real number) times the distance to the closest point seen so far, then the other leaf is visited recursively. Then the distance from the query point and the box is computed exactly using incremental distance updates. Each query point is assigned a label according to the weighted distribution of k nearest neighbors (based on Euclidean distances).

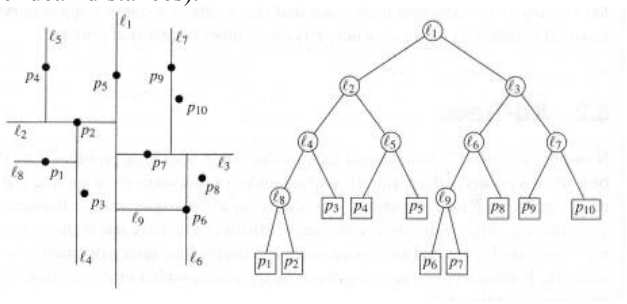

Figure 2: The organization of the sift descriptors in the kd-tree.

Detection. The query SIFT features are relocated based on the distance and orientation obtained from the nearest neighbors. Next we use the Scale Co-occurrence method for the detection of the clustered centers. In this method we create a co-occurrence matrix (N x N), where N is the total number of SIFT descriptors in the test image. If the center of a SIFT feature lies within the scale of another (Fig. 3), then the corresponding element of the row is incremented by 1. If a given feature contains a number of other features within its scale which are greater than the input threshold, a bounding box is drawn around it. The overlap of all such bounding boxes gives all the clusters in the test image. Fig. 4 shows the processing done on some test images. The middle segment of the Fig. 4 shows how clusters are formed and the right segment shows the bounding boxes of the clusters.

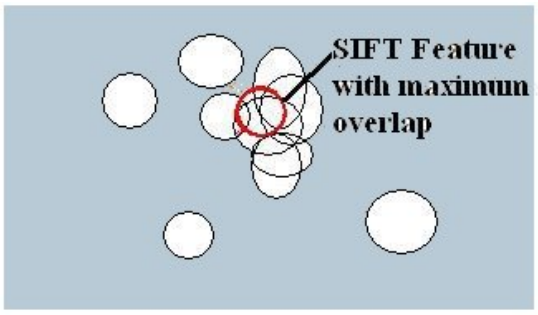

Figure 3: Scale Co occurrence Method. Overlapping regions from the scale of SIFT features are clustered into a collection of regions.

Recognition. The recognition of the traffic signs from the test image is based on the voting from SIFT features present in the regions of interest. The regions of interest span over the original positions of the SIFT keypoints which lie in the detected clusters. Once the regions of interest are defined, the SIFT features present in these regions of interest are further monitored. It is observed that the scales of the SIFT features

of a particular traffic sign are approximately of the same size. In order to prevent false votes, a tolerance is set on the scale of these SIFT features. During the calculation of the co-occurrence matrix only those features are considered whose scales are approximately equal and lie within the tolerance limit. Only these features are considered for voting. As mentioned earlier, all the trained images in the database are associated with labels (object hypotheses). Each SIFT descriptor of the test image will vote for a particular label depending on the nearest neighbors from the kd-tree. The results of all the votes are plotted on a histogram (Fig.5) .The labels which correspond to the highest peaks are selected hypotheses of traffic signs in the test image.

Time Complexity Analysis. The time required by the process is dependent on the various processing stages. The kd-tree performs with order O(N x D) where N is the number of training prototypes composed of D features. The complexity of SIFT feature extraction is dependent on the image size and performs in constant time in practice. The clustering and the voting time ranges in micro-seconds with a 1.7 GHz processor. The methodology is still in the stage of conceptual work but there is the potential to optimise the coding towards real-time processing of video frames. In particular, pre-processing the input imagery with color specific regions of interest will reduce the computational complexity to only perform highly accurate hypothesis testing with the purpose to gain reliable results.

Figure 4: (Left) The test images. (Middle) Clustering of SIFT descriptors at the centre of traffic signs. (Right) Bounding boxes of traffic signs.

III. EXPERIMENTS

The steps followed in the detection and recognition is illustrated in Fig. 6. The first step is to extract the relocated Scale Invariant features of the test image. Next we do the clustering of these features in the image. The regions of interest span over the original positions of the SIFT features. The features in these regions of interest are considered for voting. The histogram denotes the results of the voting, the peaks in the histogram represent the traffic symbols present in the image. In this section we briefly explain the configuration of the IMSERV Database and the training samples. For training, we have taken 60 samples of different traffic signs of prohibitory and warning categories. At the present we have taken only one example of each sign but the idea is to have multiple trained images for each traffic sign for improving the recognition rate. Next we state some observations about the thresholding and tolerance.

Configuration of IMSERV Database. The IMSERV database consists of the trained images and test images. This database is built up with real world images of Austrian traffic signs. Fig. 7 shows sample images which are used for detection and recognition. For the purpose of testing, images of 3 Mega-pixel resolution are captured by a static digital camera; in addition, some images are extracted from video frames of a video camera attached to a car (Fig. 8). Video frames are extracted for the purpose of testing the traffic sign classifier whereas single images from the static camera are used for training purpose. Video frames provide various scales and sizes for the testing purpose. The video taken is from the inner and outer city of Graz, Austria. There are about five sample test images for every traffic sign in the database.

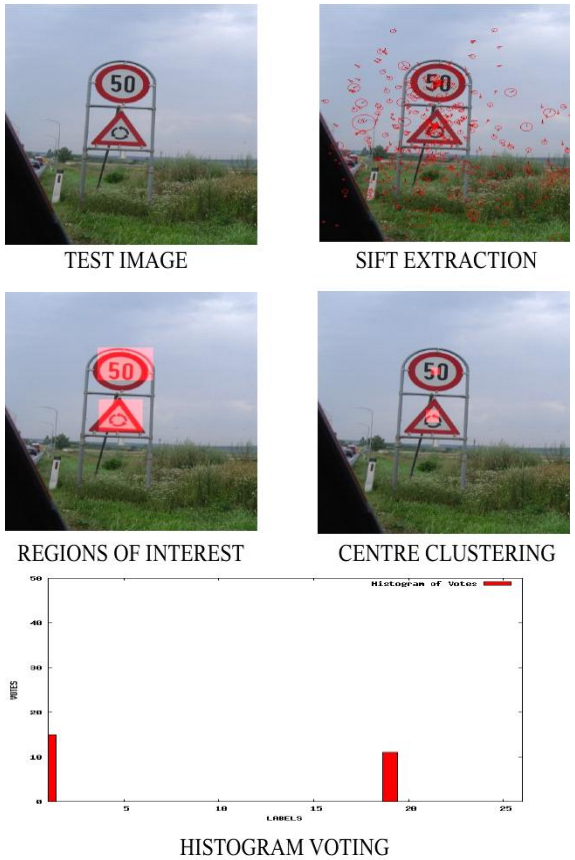

Figure 5: Individual processing stages and corresponding results in the test image.

Thresholding. As mentioned earlier (Sec. II) the detection of the traffic signs from the test image is contingent

on the threshold value. The threshold value determines the list of features that overlap with other features which in turn determine the center clusters and the regions of interest. It has been observed that for good images in which the traffic signs are prominent, a high threshold (4-5) would give a recognition rate ≈ 100% but for images in which partial occlusions or traffic signs are of a smaller scale, a low threshold (2-3) helps in the detection of the clustered centers. A detailed analysis is given in the Sec. IV.

Tolerance. The tolerance of the scale size also plays a crucial role in the recognition. Fig. 9 shows the difference in the histogram votes for the same test image. The scales of the descriptors for the traffic sign belong almost to the same size. While making the bounding boxes for the SIFT descriptors, we check if the size of the scales fall within the tolerance limit, thus preventing false votes from voting. The experiments were carried out on various images with traffic signs with different scales and affine projections. The results are discussed in Sec. IV.

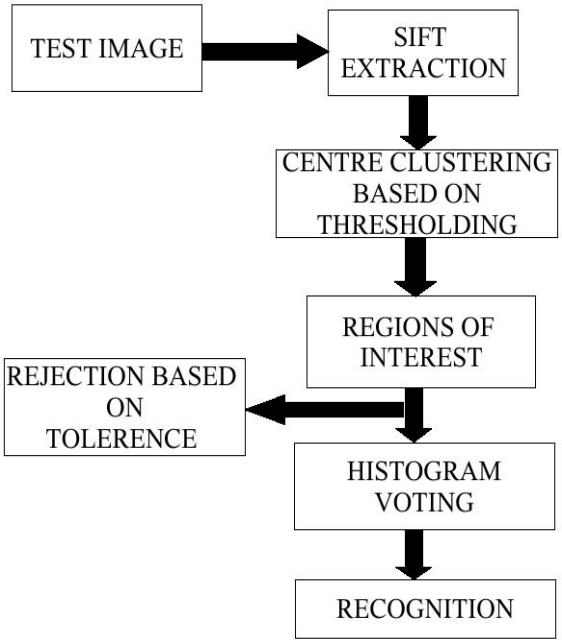

Figure 6: Schematic of the traffic sign recognition on a test image.

Figure 7: Training images from the IMSERV database.

IV. RESULTS

In this section we present and discuss the results obtained on the test images from the IMSERV database. We have considered over fifty images for the testing purpose which include the static images and the images obtained from the video. For the testing purpose we have only considered the prohibitory and the warning signs.

Invariance to Affine Projections. One of the key observations is that the recognition is rather invariant to the affine projections of the traffic sign in the image. This is mainly credited to the use of Scale Invariant Features used in the recognition. Since the descriptor is computed with respect to the dominant orientation, it is rotation invariant. Fig. 10 shows some samples of correct recognition in case of affine projections.

Low Threshold Values. The recognition of the test images is based on the threshold input. One of the problems faced by this approach is the size of the traffic sign in the image. As the size decreases, the traffic sign becomes insignificant to other objects present in the image due to which the number of features for the traffic signs decreases. The clustering at the center of the traffic sign becomes difficult in this case as features for the other objects being larger in size, there is a fair chance of false clusters being formed. For the detection of the traffic sign in this case, we need to lower the threshold. Lower threshold (2-3) implies that we relax our criteria for the cluster detection (Fig. 11). Lower thresholds are also used for images in which distorted or rusted traffic signs are present. The recognition results in case of low threshold values are ≈ 65% accurate

For the images which have prominent traffic signs (Fig. 12) high threshold values are used. The recognition in these cases is ≈ 100%. accurate..

Figure 8: Images from the IMSERV database extracted from the video frames.

Video Images. The video is taken with a cam recorder mounted on a moving vehicle. For the images obtained from the video, the recognition rate is not high, as the images have a motion blur due to vehicle vibration. As we can see in Fig. 12 the region of interest extends beyond the traffic sign. Some features are detected for the noise in the image.

Partial Invariance to Illumination. The approach is partially invariant to the illumination of light i.e. it does not depend on the time of day. The results from testing show that recognition works on images captured during different times of the day. Fig. 13 shows some of the test results. Sufficient illumination is required during night.

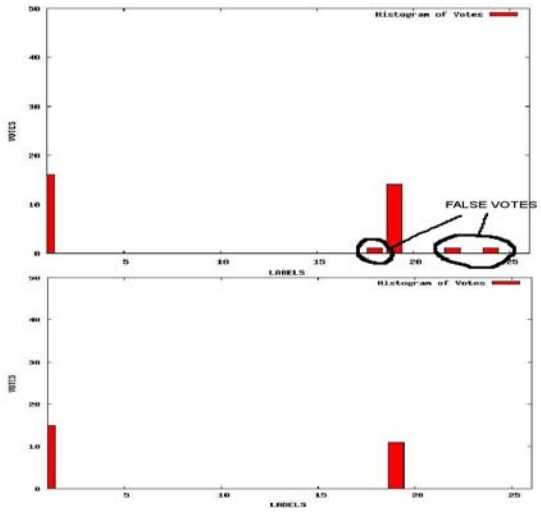

Figure 9: The histograms of votes before (above) and after imposing the tolerance limit (below).

One of the important observations in this approach is that the regions of interest obtained after applying an appropriate threshold (high or low) contain all the traffic signs in the test image i.e. the detection rate is 100%. The ambiguous voting (Fig. 11) for the lower threshold values may arise due to the features present in areas that are not including the traffic sign. To improve the recognition results we plan to implement standard K-means clustering on the descriptors as explained in Sec. V. This would bring about additional category recognition. The shapes of the traffic signs chosen for detection are mainly circular or triangular. As mentioned in Sec. II we obtain more Scale Invariant Features on the edges of the traffic sign which mainly depict the edge features (shape). Since these features are more in number, the probability of these features forming clusters is more. This is one of the factors why we propose to implement K-means clustering.

Figure 10: Recognition samples in the case of affine projections.

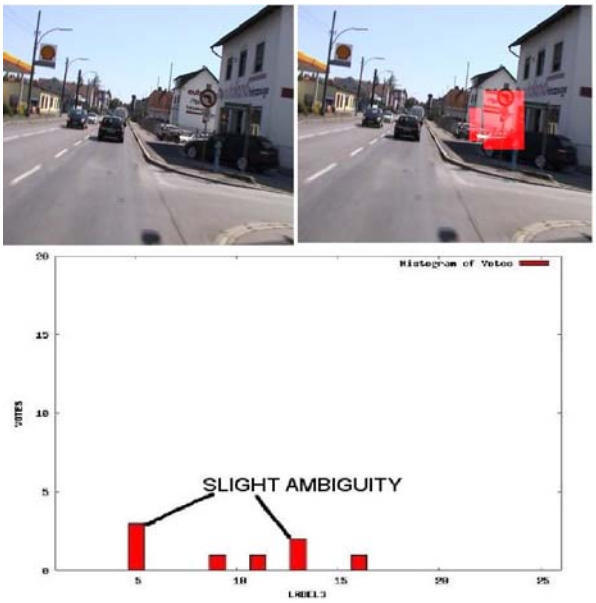

Figure 11: Detection on a video image with Low threshold (=2).Though the peak denotes the right result, there is a slight ambiguity as the second peak also has comparable number or votes.

V. CONCLUSION AND FUTURE WORK

In this paper we present an approach for traffic sign recognition based on grouping descriptors by the center clustering of the Scale Invariant Features (SIFT). The performance proves to be effective as compared to the results obtained from the unaltered voting of the Scale Invariant features. We studied the distinctive nature of Scale Invariant Features on the traffic signs and observed a high degree of invariance to affine projections. The test results prove that recognition works effectively in cases where traffic signs are of comparable size but it becomes ambiguous in the cases in which the size of the traffic sign is rather small in the image. Overall the approach is independent of color, shape, partially to illumination effects and is comparable to the state of art approaches based on powerful machine learning techniques.

Figure 12: Detection with high threshold (=5).

We are planning experiments that aim at discriminating descriptors that particularly vote for categorical information (common to all prohibitory signs, etc.) and for individual sign information (referring to the central pictorial content of a sign). Since we have observed that the features which vote for the shape of the traffic sign are more in number, they can be easily clustered for e.g., by using a k-means approach. For its implementation, we first plan to reduce the 128 dimensions of the descriptors to 40 using Principle Component Analysis. The dimensionality reduction helps clustering and also increases the efficiency. The expectation is to obtain two clusters which consist of the circular shape and the triangular shape traffic signs, respectively. The test descriptors will be further screened based on their distance from the center of the clusters in the K-means. Only those descriptors which are close to the shape clusters will be considered for voting, thus eliminating more of false votes and improving further the recognition rate.

Figure 13: Sample traffic signs and correct detection and classification under different illumination conditions.

ACKNOWLEDGMENT

This work is supported in part by the European Commission funded project MOBVIS under grant number FP6-511051 and by the FWF Austrian Joint Research Project Cognitive Vision under sub-project S9104-N13.

REFERENCES

[1] D. Lowe. Object recognition from local scale invariant features. In Proc. International Conference on Computer Vision, pages 1150–1157,1999.

[2] L. Priese, V. Rehrmann, R. Schian, and R. Lakmann.Traffic sign recognition based on color image evaluation, 1993.

[3] Arturo de la Escalera, Jose M. Armingol, and Mario Mata. Traffic sign recognition and analysis for intelligent vehicles. Image Vision Comput., 21(3):247–258, 2003.

[4] X. W. Gao, L. Podladchikova, D. Shaposhnikov, K. Hong, and N. Shevtsova. Recognition of traffic signs based on their colour and shape features extracted using human vision models. J. Visual Communication and Image Representation, 17(4):675–685, 2006.

[5] J.T. Oh, H.W. Kwak, Y.H. Sohn, and W.H. Kim. Segmentation and recognition of traffic signs using shape information. pages 519–526, 2005.

[6] D. M. Gavrila. Multi-feature hierarchical template matching using distance transforms. In International Conference on Pattern Recognition, pages 439–444, 1998.

[7] C. Y. Fang, C. S. Fuh, P. S. Yen, S. Cherng, and S. W. Chen. An automatic road sign recognition system based on a computational model of human recognition processing. Comput. Vis. Image Underst., 96(2):237–268, 2004.

[8] J.A. Janet, M.W. White, T.A. Chase, R.C. Luo, and J.C. Sutto. Pattern analysis for autonomous vehicles with the region- and feature-based neural network: global self-localization and traffic sign recognition. pages IV: 3598–3604, 1996

[9] D. Lowe. Distinctive image features from scale invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004.

[10] K. Mikolajczyk, B. Leibe, , and B. Schiele. Multiple object class detection with a generative Model. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), June 2006.

[11] Gerald Fritz, Christin Seifert, and Lucas Paletta. A mobile vision system for urban detection with informative local descriptors. icvs, 0:30, 2006.

[12] Bastian Leibe, Edgar Seemann, and Bernt Schiele. Pedestrian detection in crowded scenes.In CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) – Volume 1, pages 878–885, Washington, DC, USA, 2005. IEEE Computer Society.

[13] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 264–271, Madison, Wisconsin, June 2003.

[14] Shivani Agarwal, Aatif Awan, and Dan Roth. Learning to detect objects in images via a sparse, part-based representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(11):1475–1490, 2004.