Hazardous sign detection for safety applications in traffic monitoring

Wanda Benesova and Michal Kottman *Slovak Univ. of Technology (Slovakia)*

Oliver Sidla *SLR Engineering OG (Austria)*

Proc. SPIE 8301, 830109 (2012);
http://dx.doi.org/10.1117/12.905813

See the paper on the next page ...

Hazardous sign detection for safety applications in traffic monitoring

Wanda Benesova*[a], Michal Kottman[a], Oliver Sidla[b]

[a]Faculty of Informatics and Information Technologies, Slovak University of Technology, Bratislava, Slovakia;
[b]SLR Engineering OG, Graz, Austria

ABSTRACT

The transportation of hazardous goods in public streets systems can pose severe safety threats in case of accidents. One of the solutions for these problems is an automatic detection and registration of vehicles which are marked with dangerous goods signs. We present a prototype system which can detect a trained set of signs in high resolution images under real-world conditions. This paper compares two different methods for the detection: *bag of visual words* (BoW) procedure and our approach presented as *pairs of visual words with Hough voting*. The results of an extended series of experiments are provided in this paper. The experiments show that the size of visual vocabulary is crucial and can significantly affect the recognition success rate. Different code-book sizes have been evaluated for this detection task. The best result of the first method BoW was 67% successfully recognized hazardous signs, whereas the second method proposed in this paper - pairs of visual words and Hough voting - reached 94% of correctly detected signs. The experiments are designed to verify the usability of the two proposed approaches in a real-world scenario.

Keywords: Object detection, Local descriptors, SIFT, Visual words, Bag of visual words, Hough voting

1. INTRODUCTION

Transportation vehicles with hazardous materials can cause several dangerous situations when they exceed the traffic regulation. Hence, hazardous sign detection in traffic monitoring is a important problem from the point of view of security. Visual automated detection is therefore a challenge for computer vision researchers. Similar object detection tasks were traditionally solved by methods based on segmentation or algorithms derived from sliding window, both in combination with features detection and classification.[1] Novel techniques which used local descriptors like SIFT,[2] SURF[3] etc. are very promising in the task of object detection and recognition due to object pose, scale and rotation invariance and therefore they focus the attention of computer vision experts in the last years. We have already presented[4] an evaluation of several combinations of various key-point detection and description methods in order to find the best and reasonable combination of them. For all experiments presented in this paper, we have chosen the SIFT feature detector and feature descriptor, because of their known best accuracy despite the time consuming computation. The SURF feature detector and descriptor combination, which was evaluated in the paper[4] as a favorite combination, could be a good replacement of the SIFT in the optimization phase from the performance point of view.

2. DATASET

For the experiments presented in this paper, a ground truth dataset consisting of 186 images with resolution of 1920x1080 pixels has been used. Corresponding reference data to all images was created manually. All images capture traffic situations with cluttered background and most of them include one or more hazardous sign mounted on a vehicle.

E-mail: benesova@fiit.stuba.sk, Web: http://vgg.fiit.stuba.sk/

3. PREPROCESSING

Images used in our dataset pose a computational challenge, because the average number of SIFT key-points detected in an image exceed 10,000 without any preprocessing. Therefore we have decided to carry out two preprocessing steps: an edge-preserving smoothing technique to reduce the noise and a technique for detection of regions of interest (ROI) to reduce the evaluated area.

The first step of preprocessing was made by smoothing of the local maxima using *morphological grayscale reconstruction*[5] where the amount of removed maxima is controlled by a *mask* image. This *mask* image is created as a result of subtraction of a constant parameter from the input *marker* image. The same algorithm can also remove local minima, which become local maxima after image inversion. We have tested the values of subtractive constant of 8, 16 and 24 respective.

The second step of preprocessing aims to reduce the image area for further investigation by detecting regions of interest (ROIs) using the Maximally Stable Extremal Regions (MSER)[6] detector. The number of regions can be further reduced by exploiting the structure of hazardous signs. We therefore accept only regions whose side ratio and size are in tolerance range.

The smoothing (subtractive constant $c = 8$) has reduced the average amount of the SIFT key-points by 11% compared to the un-preprocessed data, by 22% for $c = 16$ and by 32% for $c = 24$. Furthermore, area restriction by MSER detector leads to a reduction of the number of active key-points to only 32%. Described preprocessing technique was used in all presented experiments; used subtractive constant was equal to 8.

4. CODE-BOOK OF VISUAL WORDS

Visual words are base elements used to describe an image and can be derived by clustering in the feature space. The algorithm published by Leibe et al.[7] uses Euclidean distance in the feature space for the clustering procedure. Additionally, we have reduced the dimensionality of the SIFT vector feature space from 128 to 36 by means of Principal Component Analysis (PCA). Approximately 600,000 samples of SIFT features selected from the whole dataset have then been used as an input of the k-means clustering algorithm. Our experiments show that the number of clusters is crucial and can significantly affect the recognition success rate of the created code-book. Therefore 5 different code-books with different number of clusters have been tested in our experiments: 200, 500, 1000, 2000, and 5000 words.

The algorithm of code-book generation can be described as follows:

1. Preprocessing (edge-preserving smoothing and detection of MSER)
2. SIFT key-point detection on all datasets images and descriptor calculation (approx. 600,000 in our dataset)
3. PCA for dimensionality reduction
4. k-means clustering

The cluster centers of the N clusters now form a visual dictionary, later referred to as *code-book N*. Some examples of image patches corresponding to 4 chosen visual words included in the code-book 1000 are presented in Figure 1.

In following sections, we will refer to *relevant* visual words as those visual words that occur on any of the templates of a hazardous sign. The number of relevant visual words (shown in Figure 2) is growing in the code-books 200–1000 (190/200, 341/500, 425/1000), but further the number of relevant visual words in the code-books 1000–5000 is saturated and even slightly decreasing (425/1000, 421/2000, 415/5000).

5. DETECTION OF HAZARDOUS SIGNS USING VISUAL VOCABULARY

Some visualized examples of selected features are presented in Figure 3. Features belonging to the *relevant* visual word are visualized in green color and the remaining features in red color. A direct ability to select between the detected object - a hazardous sign and a background, seems to be quite poor using the code-book 500. Code-book 1000 does not produce significantly better results in this visualization; code-book 5000 finally shows promising results.

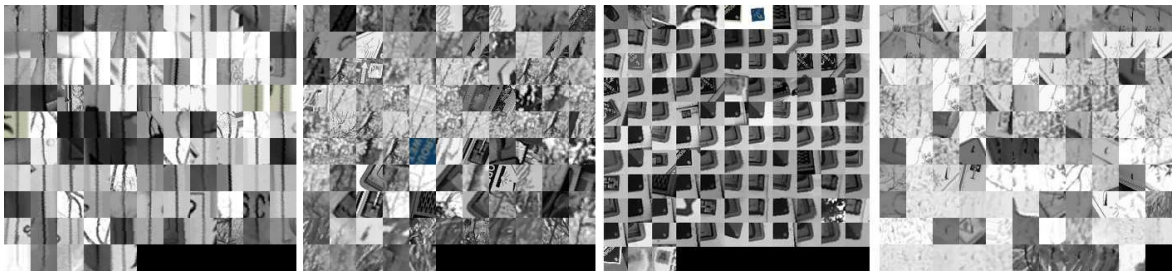

Figure 1. Examples of image patches corresponding to 4 chosen visual words.

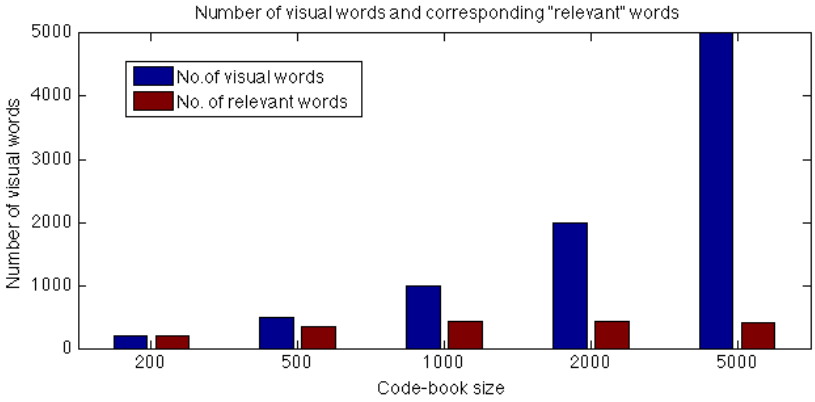

Figure 2. The number of relevant visual words based on code-book size.

6. BAG OF VISUAL WORDS

This approach is based on the method known as *bag of words* (BoW) in the natural language processing domain. The whole image can be described as a *bag of visual words*, which are determined using the detected key-points and the pre-computed code-book. The main advantages of the method are computational simplicity, efficiency, and intrinsically invariance. On the other hand, a general disadvantage of BoW is that it ignores the spatial relationships among the patches. Another disadvantage is that the BoW is computed for entire image, and the hazardous sign may get lost in the word histogram. We overcome this problem by applying BoW matching to multiple MSER regions of interest independently. This increases the possibility of detection and also allows us to determine the spatial position of the sign in an image.

The main steps of BoW method tested in our experiments are:

1. Preprocessing – edge-preserving smoothing and detection of MSER – *described in Section 3*
2. Code-book generation - *described in Section 4*
3. For all MSER regions - construct a histogram of visual words contained in the region
4. Compare all the MSER region histograms with each of sign histograms using the intersection method: $d(H_1, H_2) = \sum_{k=1}^{N} \min(H_1(k), H_2(k))$, where H_1 – template histogram, H_2 – region histogram, N – number of words
5. If $d(H1, H2) >$ threshold $\Rightarrow$ the corresponding sign is detected in the region

Four of generated code-books (size of 500, 1000, 2000 and 5000 visual words) have been evaluated using the whole dataset. A typical example of histogram intersection results is presented in Figure 4. Only one hazardous sign (template index = 9) is present on this evaluated example picture. Histogram intersections are plotted for

Figure 3. Examples of relevant features (green) and remaining features (red) using varying code-book sizes.

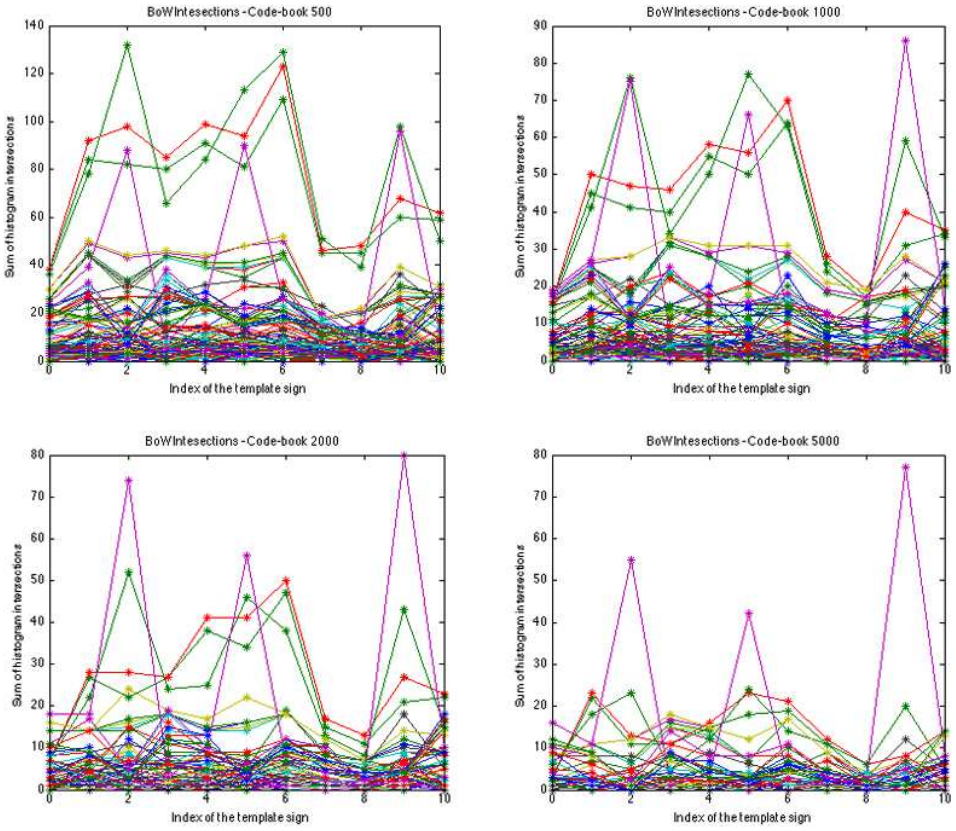

Figure 4. Plot of histogram intersections using different code-book sizes.

all sign templates (x-axis) and for all ROIs detected by MSER (one curve for each ROI). A correct detection in form of a dominant peak can be seen for sign with index 9 using the code-book 5000. In the case of code-book 500 the recognition fails, by using the code-books above 1000 the results are ambiguous. A growing number of visual words in code-book leads to better recognition results.

7. PAIRS OF VISUAL WORDS AND HOUGH VOTING

The goal of our approach is to improve the detection ability of relevant visual words. This approach is based on the idea to combine the features in pairs with their geometrical arrangement given by their original arrangement in the template. In training phase, we calculate two parameters, which are used to define the geometrical arrangement of the feature pair: α – angle difference between the two features; d – distance between the features. These parameters are stored for each pair of words for further calculation (Figure 5).

In detection phase, we select the most probable scale and rotation of the detected sign using the generalized Hough transform. The Hough space is two dimensional; one dimension is the angle difference and the second

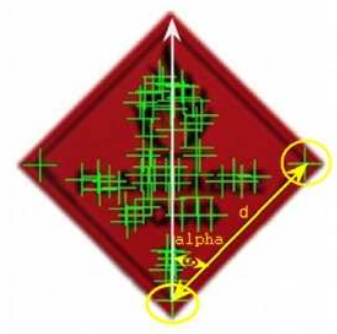

Figure 5. Pairs of words — α – angle between the two features and d – distance between the two features

dimension is the scale factor: $\alpha_{hough} = \alpha_{actual} - \alpha_{nominal}$, $d_{hough} = d_{actual}/d_{nominal}$.

The parameters of relevant feature pairs placed inside of a hazardous sign are statistically similar and are expected to create a maximum in the Hough accumulator. The parameters of random pairs of features (false detection) are expected to be statistical scattered. After the voting, we can see a maximum in the Hough accumulator on the position corresponding to the detected rotation α_{hough} and scale factor d_{hough}. Visualization of the Hough accumulator for two images can be seen in Figure 6b.

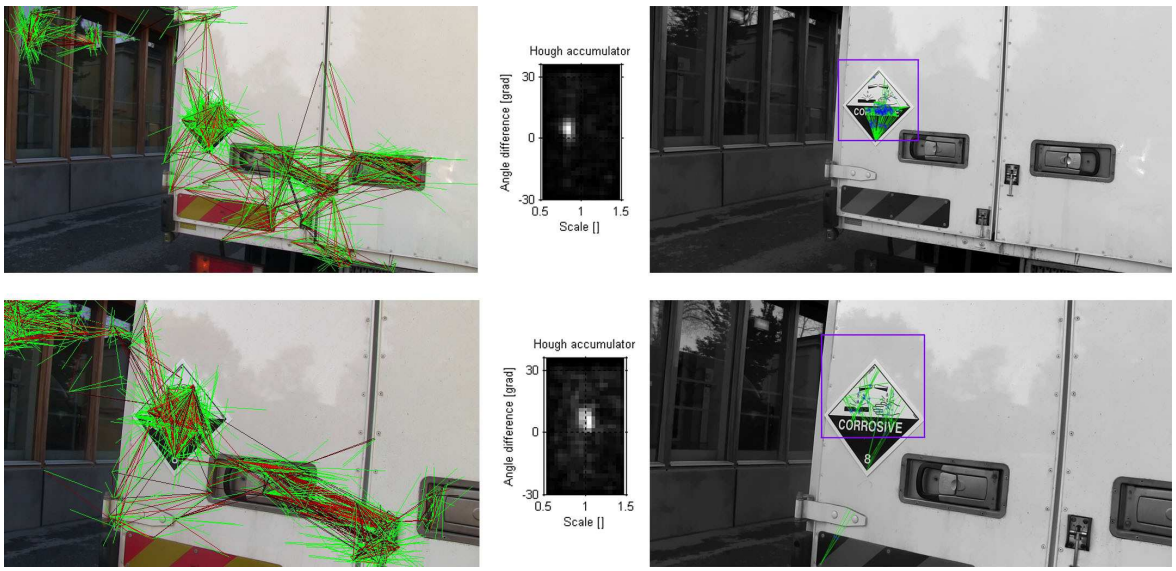

Figure 6. Hough voting for two images with varying sign scale. a)Input relevant feature pairs (green), b) Hough accumulator , c) Output feature pairs

From the results of Hough voting, we can reconstruct the image position of the pairs which have contributed to the selected maximum in the Hough space. Two examples with different size of detected sign (different d_{hough}) can be seen in Figure 6. Green lines in the figure are connection of two relevant features. All of them vote in Hough space and the voting results are shown in Figure 6b. Backprojection of the Hough accumulator maximum into the image is presented on the Figure 6c.

Once the correct rotation and scale of the sign has been determined, the results are further supplemented by spatial detection in the image. A sliding window with the size derived from the maximum in the Hough accumulator and with a large shifting offset ($windowSize/3$) counts the number of covered spatial related pairs. Centers of lines formed by the pairs are counted inside the sliding window (blue points in Figure 6c). If the

number of points inside the sliding window exceeds a given threshold t (evaluated at $t = 2$), the window is declared as a positive detection (violet rectangle in Figure 6c).

8. RESULTS

Both presented methods (*bag of visual words* and *pairs of visual words and Hough voting*) have been evaluated in the task of detection of hazardous signs using the described ground truth dataset. In addition, the influence of the code-book size (500, 1000, 2000, 5000) on the detection success has been investigated and evaluated in the BoW method. Our approach presented as *pairs of words and Hough voting* has been evaluated using the code-book 5000. The result of this method — 94% correctly detected hazardous signs — is very promising, although the percentage of false positive detection is also high, about 80%. It will be necessary to design a robust verification method to remove the false positive detections in the next stage.

Table 1. The evaluation of correct detections.

Correct detections per method / code-book size	500	1000	2000	5000
Bag of Visual Words	11%	30%	45%	67%
Pairs of Visual Words with Hough Voting				94%

9. CONCLUSION

Two methods based on visual words were presented and experimentally evaluated. The best result of the BoW method was 67% successfully recognized hazardous signs, whereas the method proposed in this paper — *pairs of visual words and Hough voting* — reached 94% of correctly detected signs. The size and quality of the code-book used is a crucial factor for the both methods. In the future work, the quality of the code-book will be investigated and optimized and a robust verification method will be designed.

ACKNOWLEDGMENTS

This work was supported by the grant KEGA 244-022STU-4/2010.

REFERENCES

[1] Benesova, W., Lypetskyy, Y., Andreu, J.-P., Paletta, L., Jeitler, A., and Hoedl, E., "A mobile system for vision based road sign inventory," in [*Proceedings of 5th International Symposium on Mobile Mapping Technology*], (2007).

[2] Lowe, D., "Distinctive image features from scale-invariant keypoints," *International journal of computer vision* **60**(2), 91–110 (2004).

[3] Bay, H., Tuytelaars, T., and Van Gool, L., "Surf: Speeded up robust features," *Computer Vision - ECCV 2006* **3951**, 404–417 (2006).

[4] Sidla, O., Kottman, M., and Benesova, W., "Real-time pose invariant logo and pattern detection," in [*Proceedings of SPIE - The International Society for Optical Engineering*], **7878** (2011).

[5] Vincent, L., "Morphological grayscale reconstruction in image analysis: applications and efficient algorithms.," *IEEE Transactions on Image Processing* **2**(2), 176–201 (1993).

[6] Matas, J., Chum, O., Urban, M., and Pajdla, T., "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing* **22**, 761–767 (Sept. 2004).

[7] Leibe, B., Leonardis, A., and Schiele, B., "Robust Object Detection with Interleaved Categorization and Segmentation," *International Journal of Computer Vision* **77**, 259–289 (Nov. 2007).