ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0011

Text Sign Detection Method based on Dual Feature Fusion

Haoyue Lu

Sias University Zhengzhou, Henan; Pengfei Song Sias University Zhengzhou, Zhengzhou, China

Abstract

Text sign detection in natural scenes faces the challenge of complex background and various colors, which imposes a great burden on irregular text recognition. Therefore, special classification and detection methods are needed. This paper proposes a text sign detection method based on dual feature fusion, which improves the detection accuracy by fusing contour and color features. In the aspect of contour features, the spatial pyramid matching model is combined with SIFT feature extraction technology, and the self-growing neural network is used to adaptively determine the number of feature categories. In terms of color processing, the HSV color space is improved. By quantizing the hue and saturation components and sorting the color distribution, the color features that are robust to illumination changes are obtained. The most important innovation is to concatenate the contour histogram and the improved color histogram to form a comprehensive feature representation with dual feature fusion. The experimental results show that the proposed method performs well in the test containing thousands of street view images, with an accuracy of 89.16% for positive samples and 94.3% for negative samples, which is significantly better than the method using a single feature. This result verifies that considering both shape and color information can better recognize text signs, which provides an effective technical solution for practical applications.

Keywords

Text Sign Detection; Irregular Text; Dual Feature Fusion.

1. Introduction

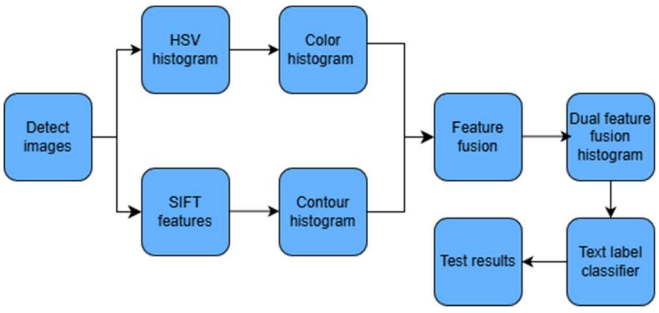

Figure 1. Text label detection process based on dual feature fusion

There are many road signs, billboards and traffic signs in natural scenes, and in real life, not every sign contains only text. The images in street view are diverse, and if they are not distinguished, it is a heavy task for the recognition of irregular text [1]. Natural scene text panels with strong texture, strong marginal characteristics, the background color of the sign has obvious different, so in order to accurately describe text panels under the natural scene, the outline of the paper combines the image text and color features, to test the word sign [2], and put forward the word sign detection method

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0011

based on double feature fusion, It consists of two steps: histogram feature extraction based on contour and color feature fusion, classifier training and detection. The detection process is shown in Figure 1.

2. Outline Features of Text Label

The traditional Bag of Visual Word (BOVW) model mainly converts the image into a set of local features that ignore spatial information and do not consider the order, and then trains and detects it by SVM trainer. In the training module, SIFT (Scale Invariant Feature Transform) operator is used to extract local features[3]. This operator can produce strong robustness when the image is rotated or contracted greatly, so it can accurately describe the image features. However, the use of SIFT operator will lose the spatial relationship of the features in the image [4]. To solve this problem, this paper proposes a spatial Pyramid (SPM) based bag of visual words model algorithm, which is used as the contour feature of the text sign. Experiments show that the algorithm can fully express the local features and global features of the image, which become the contour feature of the text sign as a whole.

2.1 SPM Representation of Features

Traditional visual word bag model[5] although made more progress in the image detection, as BOVM did not fully consider the local characteristics of image spatial relationships, leads to lack of part of the space visual feature information, in order to solve this problem, Agarw[6] will be put forward by simple relationship into visual words in pairs, Leibe[7] proposes an implicit shape model, which uses a loose star structure to describe the various shape relationships of objects in an image. Fergus[8] proposes a super feature, which is generated by the aggregation of visual words and contains spatial features. Later, Lazebnik[9] used the spatial Pyramid (SPM) algorithm to represent the spatial relationship of local features. Experiments show that the algorithm has good stability in image detection.

SPM full name is the Spatial Pyramid Matching, image examinations should be uniform principle is divided into several blocks, and the statistical features of each block, then the local features of joining together form a complete, this is the meaning of Spatial. In the block, multi-scale is used to change the granularity of the block to be both large and fine, and finally the pyramid is formed. Finally, Matching[10] is performed, that is, BOW (bag of words model) is used in the segmented image to make the local features correspond to the original image one-to-one. In detail, the higher the level of SPM stratification, the finer the granularity of the feature histogram, as shown in Figure 2, the image is evenly divided into several blocks, the figure from left to right is 1×1, 2×2, 4×4, and then the number of different shapes contained in each block of the image is counted, and the number of histograms contained in each layer is counted from left to right.Finally, the obtained histograms are combined to give their corresponding weights, and the weights increase from left to right, and then SPM is put into the SVM classifier for training and prediction. In the figure, the original image to be detected is used as the first layer of the model. It is evenly divided into four blocks to form the second layer of the SPM model, which is used as the sub-layer of the pyramid. Then the images of the sub-layer are repeatedly divided, and the pyramid model is finally constructed by repeated division. Therefore, SPM is constructed by dividing the image blocks exponentially by 2, that is, $2^l \times 2^l$ (= 0,1,2,... , L), which is the scale of the pyramid. The SPM model uses the bag-of-words visual model to calculate the visual words histogram feature on each image patch. Finally, the above features are concatenated to form a histogram feature vector to describe the image. Model for the SPM BOVM word bag of experiments show that the algorithm can generate has powerful description ability of feature descriptor.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0011

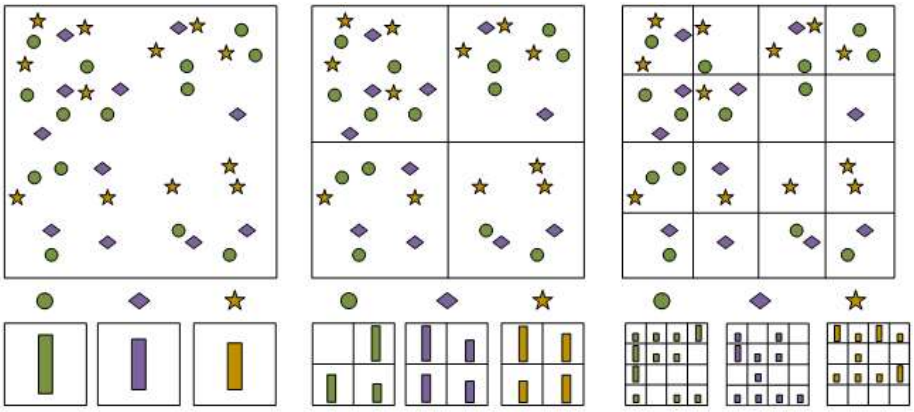

Figure 2. Spatial Pyramid Model

2.2 Generation of Contour Histogram Features

The process of contour histogram feature generation includes extracting SIFT features, constructing SPM model, and obtaining the visual dictionary to quantify SIFT features to generate contour histogram based on SGONG training. The SIFT algorithm proposed by Lowe et al.[11] in 2004 is an image matching algorithm based on point features. SIFT descriptor is invariant to rotation and scale, so it can describe image information more accurately. The calculation method is divided into four steps: constructing the image scale space, locating the key point information, determining the direction of the key point, and generating the SIFT feature vector. In order to solve the problem that people need to define the number of clusters by themselves in the traditional bag of visual words model, Atsalaki[12] proposed Self-Growing and Self-Organized Neural Gas network (SGONG). The advantage of SGONG is that it can adaptively determine the number of clusters according to the target data, which can reduce the differences within clusters and increase the differences between clusters as much as possible. In this paper, we use SGONG clustering method for BOVW dictionary learning, which can adaptively determine the number of target types in natural images, generate clustering results more accurately and reliably, and greatly reduce the complexity of image detection. The process of constructing the descriptor in SPM's bag of visual words model is as follows: the image is divided into multiple sub-image blocks to construct the pyramid model; SIFT features of each sub-image block in the pyramid model are extracted. According to the feature sub-calculation method of the bag of visual words model, the histogram features are generated. Finally, these scattered histogram features are merged to form a complete global histogram feature.

In this section, the repeated extraction of SIFT features on the spatial pyramid will lead to the increase of time complexity. However, by combining the SPM model with the location information of SIFT feature points, the SIFT feature only needs to be extracted once on the image, and the image block of each layer can be obtained by the coordinate position of the feature points. The specific steps are as follows:

Step 1: Extract the SIFT features of the image, and get the coordinates of the feature points (x, y), and the image size M_1* N_1.

Step 2: Construct the SPM model in 2-D space by dividing the image into sub-image blocks. As shown in Figure.3, the image is divided into 2^l*2^limage blocks in the space of scale, and they are labeled as p.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0011

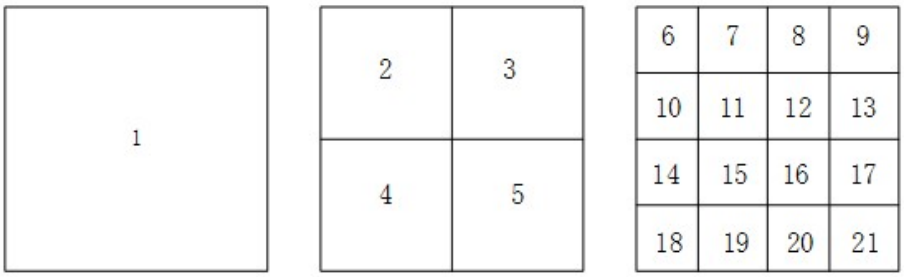

Figure 3. Schematic diagram of image labeling

Step 3: According to the image patch p obtained from SIFT feature points in Step 2, the data vocabulary histogram in each image patch is obtained based on the dictionary trained by SGONG $H\_V_p^{j_v}$(p=1, 2, …, 21, j_v=1,2,…C), C stands for the number of SGONG dictionary words. Finally, merge the histogram vectors H_V of each image block according to the formula and normalize them to obtain the final contour histogram H_S, and finally obtain the global description of the image.

2.3 Color Characteristics of Text Signs

The color feature of natural scene images is an important feature of image detection, and color is also the basic element of image composition. In natural scenes, there are blue sky, white clouds, green grass and so on. These objects reflect obvious visual characteristics. In order to facilitate people's reading, the text in an ordinary text sign is often composed of a single color, and the whole text sign is usually composed of two or three main colors. However, in natural scenes, the background of text signs and the color of the text may be rich and diverse, and the hue is not the same. Color features in image detection usually include color histogram, dominant color histogram, color moment, color set, and so on. In this paper, the HS color histogram is improved to generate the color histogram feature, which can accurately describe the relative color distribution of text signs, so as to be more beneficial to the accurate detection of text signs.

1) Calculate HS color histogram

As early as 1978, researchers proposed a color space HSV created according to the intuitive characteristics of color observed by human vision. HSV is composed of Hue, Saturation, and Value, where H and S represent color information, and V represents the brightness of the color. The HSV model reflects the intuitive understanding of color in human vision. The HSV color histogram composed of one or more components can directly reflect the image information in people's eyes, including basic hue information and color distribution. The range of hue H component is between [0,360], and the range of saturation S component is between [0,1]. In order to prevent the number of bins in the HS histogram from increasing the time complexity, it is necessary to quantize HSV first, and then count the color histogram. The specific steps are as follows: (1) According to the experimental results of human vision and color contrast, the HS is quantized, in which the H component is quantized into 16 parts, the S component is quantized into 8 parts, and 128 bins are merged. (2) The color information of the pixel feature in the image is quantized into the H and S components of the pixel respectively, and the HS color histogram is finally obtained by counting the pixels in each bin of the HS color histogram. This histogram counts the proportion of each color in the image, which represents the global feature of the image. Text signs are often composed of multiple colors, and the background colors of the signs are also different, so the generated HS color histograms are also different. However, the HS color histogram extracted from the text signs in the natural scene has many peaks. Obviously, the effect of dividing the text signs by the traditional HS color histogram is not good.

2) Invariant color histogram

In order to solve the problem of the difference of text signs caused by color, the method of maximum displacement of "HS" component in HSV color space is used to obtain an invariant color histogram,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0011

which describes the "color" in the image. Under normal circumstances, the color distribution characteristics of text signs can be used to describe the color histogram features, but the background color of text signs is messy. In order to highlight the relative distribution of the background and text color of text signs, this paper improves the HS color histogram, and the specific methods are as follows:

Step 1: The H component of the original HS color histogram is quantized into 16 parts, and the S component is quantized into 8 parts. One H component corresponds to 8 S components, and a 128-dimensional HS color histogram is obtained.

Step 2: People are most sensitive to H component in HSV color space, followed by S component and V component. Firstly, for each H component value, the color invariant histogram algorithm is used to calculate its sub-color histograms, and a total of 16 sub-color histograms are counted. Then, the values in each sub-color histogram are sorted from large to small and shifted to obtain the final color histogram sorted by H component value.

Step 3: According to each sub-color histogram, eight bin values are counted and sorted from high to low. As shown in Figure 4, the sub-color histogram is sorted for H=0. The following figure shows the color histogram features with small difference in color distribution obtained by sorting each sub-color histogram.

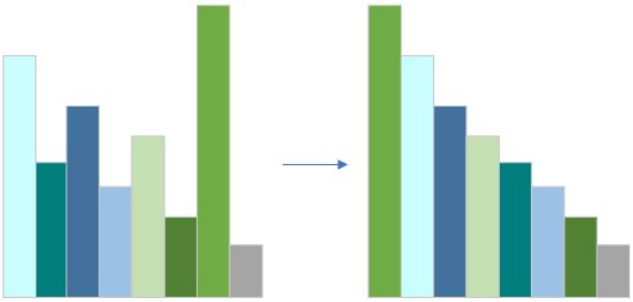

Figure 4. Sorting of sub-color histograms

2.4 Dual Feature Fusion of Text Signs

In order to fully represent the shape, color and spatial information of the image, we fuse the contour histogram feature and the invariant color histogram feature to generate the dual-feature fusion histogram feature, which expresses the overall characteristics of the image more completely. Feature fusion is generally divided into serial feature fusion and parallel feature fusion. Serial feature fusion is to combine multiple features of an image into a new feature, and then train the classifier. In this paper, the contour histogram feature and color histogram feature are combined to generate the dual feature fusion histogram feature by using serial feature fusion method. The fusion framework is shown in Figure.5. The specific steps are as follows: firstly, the contour and color features are extracted from the image, and the contour dictionary and color dictionary obtained by SGONG training are used to quantify these two features and obtain the histogram. Finally, the two feature vectors are fused to obtain the final dual-feature fusion histogram feature vector H_SC.

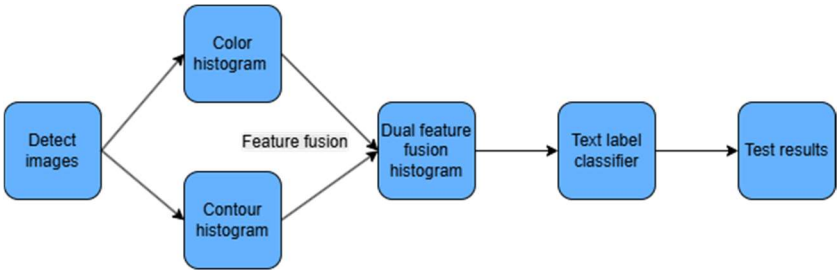

Figure 5. Fusion process of text label classifier

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0011

2.5 Training and Detection of the Classifier

Choosing an appropriate classifier for text sign detection will improve prediction accuracy, speed up computation, and enable accurate prediction models on larger data sets. The decision tree that we are familiar with is a simple classifier, which is mainly used to train problems with a small number of samples. When it comes to the classification of complex problems, it is obviously not appropriate to choose a decision tree. The Bayesian classifier assumes that each feature condition is independent and predicts the posterior probability according to the prior probability of the training sample. However, in the actual situation, the various feature attributes are closely connected, so it is not appropriate to use the Bayesian trainer in the actual situation. A large number of parameters, such as network topology, weights and thresholds, need to be initialized to train neural network classifiers. And the initial value of the threshold is often selected according to experience, and there is no clear judgment standard.

In this paper, the gradient boosting decision tree GBDT classifier is used to train the text signs. GBDT can get higher accuracy with relatively less parameter tuning time. The training process is to convert the training image into HSV color space, and generate HS color histogram to obtain the final color histogram. Then SIFT features are extracted, and the contour histogram is generated by SGONG dictionary and SPM quantization. The dual feature set is obtained by feature fusion of color histogram and contour histogram, which is trained to generate a text sign classifier. When detecting the image, the final dual-feature fusion histogram feature is obtained through this process, and then the trained text sign is used to determine whether it is a text sign.

3. Experimental Results and Analysis

3.1 Experimental Environment and Data

In order to verify the algorithm, the experiment is carried out in the environment of VC6.0. The experimental data set contains 4121 street view images of text signs. The GBDT classifier is used to classify the text sign images and street view images. In these image databases containing text signs, 500 positive samples are used for training, 1000 negative samples are used for testing, and 1500 negative samples are used for testing.

3.2 Experimental Results and Analysis

In order to verify the effectiveness of the two-feature fusion histogram algorithm proposed in this paper, we train and generate GBDT classifiers on SIFT features and contour histogram features respectively, and then perform detection. There are 500 positive samples (text signs) for training, 1000 negative samples (non-text street view images), 1000 positive samples and 1500 negative samples for testing. The results are shown in the table, where the positive and negative accuracy rates are calculated according to the formula, p is the accuracy rate, the number of positive and negative samples correctly classified, and N is the total number of positive and negative samples.

Table 1. Comparison of detection results

Feature	Positive sample accuracy	Negative sample accuracy
SIFT features	80.9%	91.93%
Contour histogram features	82.4%	91.16%
Dual feature fusion	89.16%	94.3%

Table 1 shows that in the detection results of the classifier trained by features, the accuracy of positive samples is as high as 89.16%, which is 8.26% and 6.76% higher than the accuracy trained by SIFT features and shape histogram features. The accuracy of negative samples is as high as 94.3%, which is 2.37% and 3.14% higher than that of SIFT feature and shape histogram feature training

ISSN: 2414-1895 DOI: 10.6919/ICJE.202512_11(12).0011

respectively. It can be seen that the accuracy of the proposed method is significantly improved, and it has a strong ability to describe the text signs.

4. Summary

In natural scenes, text signs always show obvious color distribution characteristics and strong texture characteristics, and the detection performance is affected by both contour and color. In this paper, the contour histogram and color histogram features are fused, and the histogram feature of dual feature fusion is proposed, which combines the contour feature and its invariant color feature. It fully describes the contour and color features of the image. In order to verify the algorithm, the GBDT classifier is used to train and detect it. The results show that the proposed dual-feature fusion algorithm significantly improves the detection rate of text signs and has good description ability.

References

[1] Fan Tairan Research on Text Detection Technology for Signage Images [D]. Jiangsu Ocean University, 2022.

[2] Lu, M.; Leng, Y.; Chen, C.-L.; Tang, Q. An Improved Differentiable Binarization Network for Natural Scene Street Sign Text Detection. Appl. Sci. 2022, 12, 12120.

[3] Agarwal H. Designing of a Resilient Watermarking Scheme Utilizing Lifting Wavelet Transform and SIFT for Videos[J]. Journal of Multimedia Processing, 2023, 12(4): 215-229. DOI:10.1234/jmp.2023. 0045.

[4] Cao Ye, Guo Lihong, Dan Jintao, etc Key point extraction of point cloud based on curve fitting and SIFT operator [J]. Industrial Control Computer, 2024, 37 (11): 45-46+49.

[5] Zheng Jinsong, Gu Haihong, Jiang Qinggang, etc Soybean leaf disease recognition based on local features and visual bag of words model [J]. China Agricultural Machinery Chemistry Journal, 2024, 45 (08): 204-209. DOI: 0.13733/j.jcam.issn.2095-5553.2024.08.29.

[6] Agarwal S, Awan A, Roth D. Learning to detect objects in images via a sparse, part based representation[J]. IEEE transactions on pattern analysis and machine intelligence, 2004, 26(11): 1475-1490.

[7] Leibe B, Leonardis A, Schiele B. Combined object categorization and segmentation with an implicit shape model[C]//Workshop on statistical learning in computer vision, ECCV. 2004, 2(5): 7.

[8] Fergus R, Fei-Fei L, Perona P, et al. Learning object categories from Google's image search[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. IEEE, 2005, 2: 1816-1823.

[9] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]//2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). IEEE, 2006, 2: 2169-2178.

[10] GUO C ,LI J ,WU W , et al.A new-generation source mechanism catalogue for historical moderate-to-strong earthquakes in the Sichuan-Yunnan region constrained by a topographic high-resolution 3D velocity model and seismic waveform matching[J/OL].Science China Earth Sciences,1-24[2025-11-16].https://link.cnki.net/urlid/11.5843.p.20251113.1040.004.

[11] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60: 91-110.

[12] Yang Yao Research on Text Sign Detection and Localization Method in Natural Scenes Based on Learning [D]. Xi'an University of Technology, 2016.