Segmentation of Graphical Objects as Maximally Stable Salient Regions

Su Yang
College of Computer Science and Technology
Fudan University
Shanghai 201203, China
E-mail: suyang@fudan.edu.cn

Yuanyuan Wang
Department of Electronic Engineering
Fudan University
Shanghai 200433, China
E-mail: yywang@fudan.edu.cn

***Abstract*—Symbol segmentation is the key to affect the performance of symbol recognition in natural scenes. As experimentally confirmed, MSER is effective in segmenting text but not applicable to segmentation of graphical objects like traffic signs. We propose a color space graphical object segmentation method. It extracts stable region of interest by applying different color similarity threshold to evaluate the stability of proximity among pixels. Experimental results show that it outperforms MSER in segmenting graphical object.**

Keywords-Image Segmentation; Symbol Segmentation; Symbol Detection; Object Detection; Region of Interest

I. INTRODUCTION

Symbol recognition in natural scenes plays an important role in a variety of applications, for instances, road sign recognition for diver assistance [1-3], license plate recognition for transportation management [4], and camera based symbol recognition for environment awareness [5,6]. There exist a couple of different schemes for symbol recognition in natural scenes: The scheme presented in [1,2] consists of 3 phases: image segmentation supervised by color signatures, filtering based on shape signatures or machine learning, and classification [1]. The method presented in [3] applies adaboost based machine learning technique to detect directly candidate blocks without segmentation at all, and perform classification in the next phase. The scheme proposed in [4] is composed of 2 phases: Detect regions of interest (ROI) in an unsupervised manner by using the method referred to as maximally stable extremal regions (MSER) and then perform shape classification. In [5], prior knowledge about edge, color, and layout of text are used for locating text regions and OCR is performed subsequently. In [6], vectorization is performed on edge points based on which a vectorial pattern descriptor can be obtained. Then, symbol spotting is performed via a voting procedure to detect the regions whose signatures are mostly consistent with reference patterns.

Since high performance can be achieved in recognizing isolated symbols even in adverse conditions by using the existing techniques such as the method proposed in [7], the critical issue for symbol recognition in natural scenes is how to segment symbols from background. Although symbol spotting is an alternative solution [6], the investigation is at very beginning and the technique is not robust yet due to the error-prone vectorization. Machine learning based symbol detection without segmentation like the method used in [3] usually results in candidate blocks, not candidate objects. This is not expected because the subsequent feature extract must be contaminated by a small portion of background contained in the candidate blocks. For supervised segmentation, it is usually based on some rules according to human observation [1,2,5]. The drawback of such methods is that human-summarized rules are usually not complete and not generic. Such rules can work only under some known or controlled circumstances but the outdoor environment is very complex and varies greatly with time and place. Due to the above reasons, we prefer to use unsupervised segmentation methods to detect regions of interest. In this track, a representative work for symbol segmentation is MSER [8], which has been widely used and explored in various applications regarding document analysis and recognition [4,9,10] and received much attention in the literature of computer vision so far [11,12]. As shown in this study, however, MSER is not applicable to segmentation of graphical objects like traffic signs since it is a gray-scale operator. Hence, we propose a new color-space segmentation method by borrowing the multi-level threshold idea from MSER. First, we compute the pair-wise similarity between image pixels in terms of color. Then, we use different threshold values to determine the proximity of every pixel for image growing. As a result, different region of interests can be obtained at multiple scales (thresholds). We then locate stable regions as change little at adjacent scales, which are referred to as stable salient regions. Finally, the maximally stable salient regions (MSSR) from all scales are selected to vote for every pixel's class label. If a pixel belongs to many MSSRs, we assign the class label of the largest MSSR containing this pixel to it.

The rest of the paper is organized as follows. We present the algorithm in section 2. In section 3, we provide the experimental results. We conclude in section 4.

II. IMAGE SEGMENTATION

The whole image segmentation procedure is composed of the following sequenced phases: (1) region grouping, (2) salient region detection at every scale, (3) stable salient region detection at every scale, (4) maximally stable salient region (MSSR) preserving, and (5) final segmentation across all scales by labeling every pixel with the class label of the largest MSSR to which this pixel belongs. In the following, we will detail every phase. A flowchart of the algorithm is shown in Fig. 1.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.46

A. Region Grouping

For a pixel, it can be represented as a vector in color space by means of RGB or HSV values. To form a region, we need to connect pixels together. Here, whether an image pixel can be connected to its neighbors is subject to the color similarity between this pixel and its neighbors. That is, only the nearby pixels with consistent color can be connected. Provided the RGB values of two given image pixels are $\boldsymbol{V}_i=[R_i,G_i,B_i]^T$ and $\boldsymbol{V}_j=[R_j,G_j,B_j]^T$, then, the color similarity between the two pixels is defined as:

$$S_{i,j}=\frac{V_i^T V_j}{\left\|V_i^T\right\|\cdot\left\|V_j\right\|} \tag{1}$$

where $\|.\|$ means norm of vector. Let $X_{N(i)}$ represent one of the 8 neighbors of image pixel X_i. Whether X_i and $X_{N(i)}$ can be connected is determined by the following equation:

$$C_{i,N(i)}=\begin{cases}0 & S_{i,N(i)}\le\theta\\ 1 & S_{i,N(i)}>\theta\end{cases} \tag{2}$$

where θ is a threshold value and $C_{i,N(i)}=1$ means that X_i and $X_{N(i)}$ can be connected. Based on Eq. (1) and Eq. (2), we can obtain the connectivity from every point to its neighbors, that is, $\{C_{i,N(i)}\}$. With $\{C_{i,N(i)}\}$, we can connect the pixels of an image to form a couple of regions. The formal definition of a region is as follows:

Definition 1 (*Region*): A region is a point set consisting of a couple of pixels, for each of which at least one of its 8 neighbors must belong to this point set.

B. Gradually Appearing Salient Regions

Note that the result of region grouping is subject to the threshold value θ defined in Eq. (2). Different region grouping results can be obtained if different threshold values are applied. So, how to set such threshold value is critical for image segmentation. Suppose that there are in total M pixels in an image. By sorting the color similarity between every two pixels, namely $\{S_{ij}|i,j=1,2,\ldots,M \text{ and } i\neq j\}$, in ascending order, we obtain a sequence denoted as $s_1\le s_2\le\ldots\le s_{M(M-1)/2}$. If we let $\theta=s_1$, then, all the pixels form a unique region because the color similarity between any pair of pixels is greater than the threshold value. If we set $\theta>s_{M(M-1)/2}$, then, there will appear M regions and every pixel forms a region. If we first set the threshold to be a small value and gradually increase the threshold value, then, an interesting phenomenon can be observed. As shown in Fig. 2, when the threshold value is not big, most pixels are connected to form a dominant area in terms of size and only a few of small regions are separated from the dominant area. The colors of such small regions must be the most distinctive ones in contrast to the dominant area. We refer to such small regions as salient regions. When the threshold value becomes bigger, increasingly more salient regions will appear. This is due to that the dominant area will fall into a couple of parts when the threshold to determine color consistence is increasing. As a result, more regions with distinctive colors in contrast to the background (dominant area) will outstand. Note that we use different colors to represent different regions in Fig. 2. Here, a formal definition of salient region and dominant region is given below:

Definition 2 (*Salient Region* and *Dominant Region*): Suppose that an image is separated into K regions $P_1,P_2,\ldots,P_K$ without any overlap between any two of them. The dominant region is defined as P_I with $I=\arg\max\{|P_i|: i=1,2,\ldots,K\}$, where $|P_i|$ means the number of pixels contained in P_i. Then, $\{P_1,P_2,\ldots,P_K\}$-P_I are defined as salient regions, that is, the salient regions are $\{P_1,P_2,\ldots,P_K\}$ excluding P_I.

C. Stable Salient Region

If the pixels composing a salient region remain mostly unchanged when the threshold value is increased to the next scale, such a salient region is referred to as a stable salient region. The formal definition is as follows:

Definition 3 (*Stable Salient Region*): Suppose that a salient region obtained at the ith scale is represented as a point set P. If there is a salient region Q at the $(i+1)$th scale satisfying $|P\cap Q|/|P|\ge\theta_R$, then, P is a stable salient region, where θ_R is a threshold close to 1.

D. Maximally Stable Salient Region

At different scales, we may obtain duplicate salient regions. If a salient region obtained at a former scale is mostly included in a salient region obtained at a latter scale, then, the former one is regarded as a part of the latter one and redundant. By removing redundant salient regions, we obtain the so-called maximally stable salient regions, the formal definition of which is presented below:

Definition 4 (*Maximally Stable Salient Region*): Suppose that P is a stable salient region obtained at the ith scale and Q represents a stable salient region obtained at the jth scale, where $j>i$. If $|P\cap Q|/|P|\ge\theta_R$ does not hold for any Q, then, P is referred to as maximally stable salient region (MSSR).

Note that if a region mentioned previously contains too few pixels, it is regarded as noise and removed.

E. MSSR Size based Pixel Labeling across All Scales

With each threshold value θ, we may obtained a couple of MSSRs, which can be used as segmented candidate objects for further processing, for instance, object recognition. For each pixel, it may belong to multiple MSSRs across different scales, that is, it may possess multiple class labels. To assign each pixel a unique class label, we summarize the MSSRs at all scales to reach a final decision for image segmentation. The detailed implementation is as follows: For every pixel, if it belongs to more than one MSSR, the final decision is that the class label of this pixel is assigned to be the same as that of the largest MSSR, which contains the largest number of pixels among those MSSRs. As such, each pixel will possess a unique class label finally.

F. Stable Salient Region with Internal Edge based Filter

Prior to final image segmentation across all scales, a filtering process can also be employed to remove the stable

salient regions that contain many edge points. The motivation is: If many edge points are include in a region, there must exist interior structures in this region. To make the interior structures salient, the region including many edge points as such should be removed. As a result, we have two solutions, one with internal edge based filter and the other without. The whole flowchart of the two solutions is illustrated in Fig. 1.

III. EXPERIMENTS

We collected images from internet to evaluate the proposed method. The images are natural scene images containing traffic signs. Our goal is to segment traffic signs from the natural scene images. Some results are shown in Fig. 3. Here, we compare our method with MSER. The five columns of Fig. 3 correspond with the raster image and the region detection results using MSSR, MSSR- (MSSR plus internal edge based filtering), MSER+, and MSER-, respectively. It is obvious that for the first 7 rows, either MSSR or MSSR- can detect the meaningful graphical objects (traffic signs) or the characters included in the graphical objects while MSER+ and MSER- miss all objects. For the 8th row, MSSR can segment both graphical objects and text but MSER can only detect a portion of the text. For the last 4 rows, MSER performs better in detecting characters but fail to detect all graphical objects. In contrast, MSSR can detect all graphical objects but fail to segment characters from the background. Note that we use different colors to represent different regions in Fig. 3. We also test MSSR and MSER with other additional images and similar to the cases shown in the first 7 rows of Fig. 3, MSSR can segment graphical objects and text from backgrounds when MSER misses all.

The common characteristic of MSER and MSSR is that both methods are based on stable region decision. The main difference between them is as follows: The key of MSER is to apply a multi-scale binarization scheme for region growing. In contrast, MSSR is actually a multi-scale clustering scheme performed in color space, which is based on color consistence. We attribute the performance difference between the two methods to their algorithmic difference as described above.

IV. CONCLUDING REMARKS

Object detection or segmentation is essential for symbol recognition in natural scenes. MSER is an excellent method for detecting text and widely used in a variety of applications like license plate recognition. However, it is not able to deal with graphical symbols in color images with complex background. This motivates us to develop a new method that can effectively segment graphical symbols from natural scenes. The experimental results show that the proposed method performs better than MSER in detecting graphical objects. The major difference between MSER and the proposed method is: MSER is based on multi-level binarization while MSSR results from multi-level clustering based on color consistence.

The proposed method can also be applied in other color spaces like HSV. As we have tested, however, the result is not comparable to that obtained in RGB space. This contradicts the widely accepted point that RGB is inferior to other color spaces for human perception. We will investigate into this issue further.

ACKNOWLEDGMENT

This work is supported by 973 Program (grant No. 2010CB731401), Natural Science Foundation of China (grant No. 61071133), Major Program of Natural Science Foundation of China (grant No. 91024011), and Science and Technology Commission of Shanghai Municipality (grant No. 09JC1401500, 08DZ2271800, and 09DZ2272800).

REFERENCES

[1] S. Maldonado Bascón, J. Acevedo Rodríguez, S. Lafuente Arroyo, A. Fernndez Caballero, F. López-Ferreras: "An optimization on pictogram identification for the road-sign recognition task using SVMs", Computer Vision and Image Understanding, vol. 114, pp. 373−383, 2010

[2] Y. Y. Nguwi and A. Z. Kouzani: "Detection and classification of road signs in natural environments", Neural Computing and Applications, vol. 17, pp. 265-289, 2008

[3] X. Baró, S. Escalera, J. Vitrià, O. Pujol, and P. Radeva: "Traffic Sign Recognition Using Evolutionary Adaboost Detection and Forest-ECOC Classification", IEEE Transactions on Intelligent Transportation Systems, vol. 10, no. 1, pp. 113-126, 2009

[4] Michael Donoser, Clemens Arth, and Horst Bischof: "Detecting, Tracking and Recognizing License Plates", ACCV 2007, Lecture Notes in Computer Science, Springer, 2007, vol. 4844, pp. 447-456, DOI: 10.1007/978-3-540-76390-1_44

[5] X. Chen, J. Yang, J. Zhang, and A. Waibel: "Automatic detection and recognition of signs from natural scenes" IEEE Transactions on Image Processing, vol. 13, no. 1, pp. 87-99, 2004

[6] M. Rusinol, J. Llados, P. Dosch: "Camera-based graphical symbol detection", 9th International Conference on Document Analysis and Recognition, IEEE Press, 2007, vol. 2, pp. 884-888, DOI: 10.1109/ICDAR.2007.4377042

[7] S. Yang: "Symbol recognition via statistical integration of pixel-level constraint histograms: A new descriptor", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 2, pp. 278-281, 2005

[8] J. Matasa, O. Chum, M. Urban, and T. Pajdl: "Robust wide-baseline stereo from maximally stable extremal regions", Image and Vision Computing, vol. 22, pp. 761–767, 2004

[9] Michael Donoser, Silke Wagner, and Horst Bischof: "Context information from search engines for document recognition", Pattern Recognition Letters, vol. 31, no. 8, pp. 750-754, 2010

[10] Reinhold Huber-Mörk, Herbert Ramoser, Harald Penz, Konrad Mayer, Dorothea Heiss-Czedik, and Andreas Vrabl: "Region based matching for print process identification", Pattern Recognition Letters, vol. 28, no. 15, pp. 2037-2045, 2007

[11] P. E. Forssen and D. G. Lowe: "Shape Descriptors for Maximally Stable Extremal Regions", IEEE 11th International Conference on Computer Vision, IEEE Press, 2007, pp. 1-8, DOI: 10.1109/ICCV.2007.4409025

[12] Y. X. Lan, R. Harvey, J. R. P. Torres: "Finding stable salient contours", Image and Vision Computing, vol. 28, no. 8, pp. 1244-1254, 2010

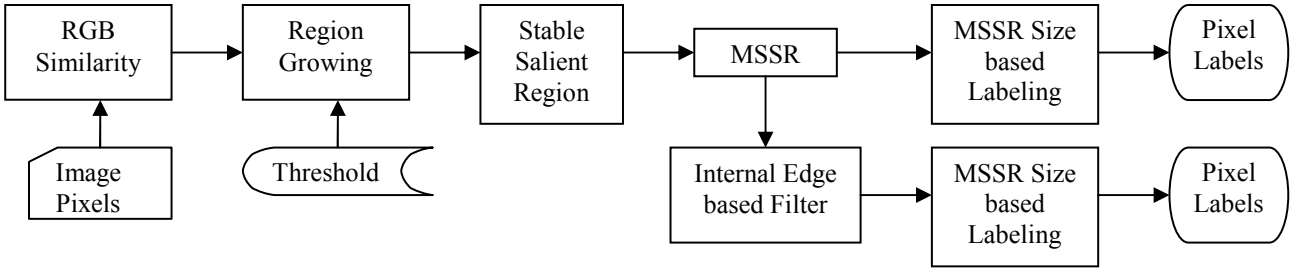
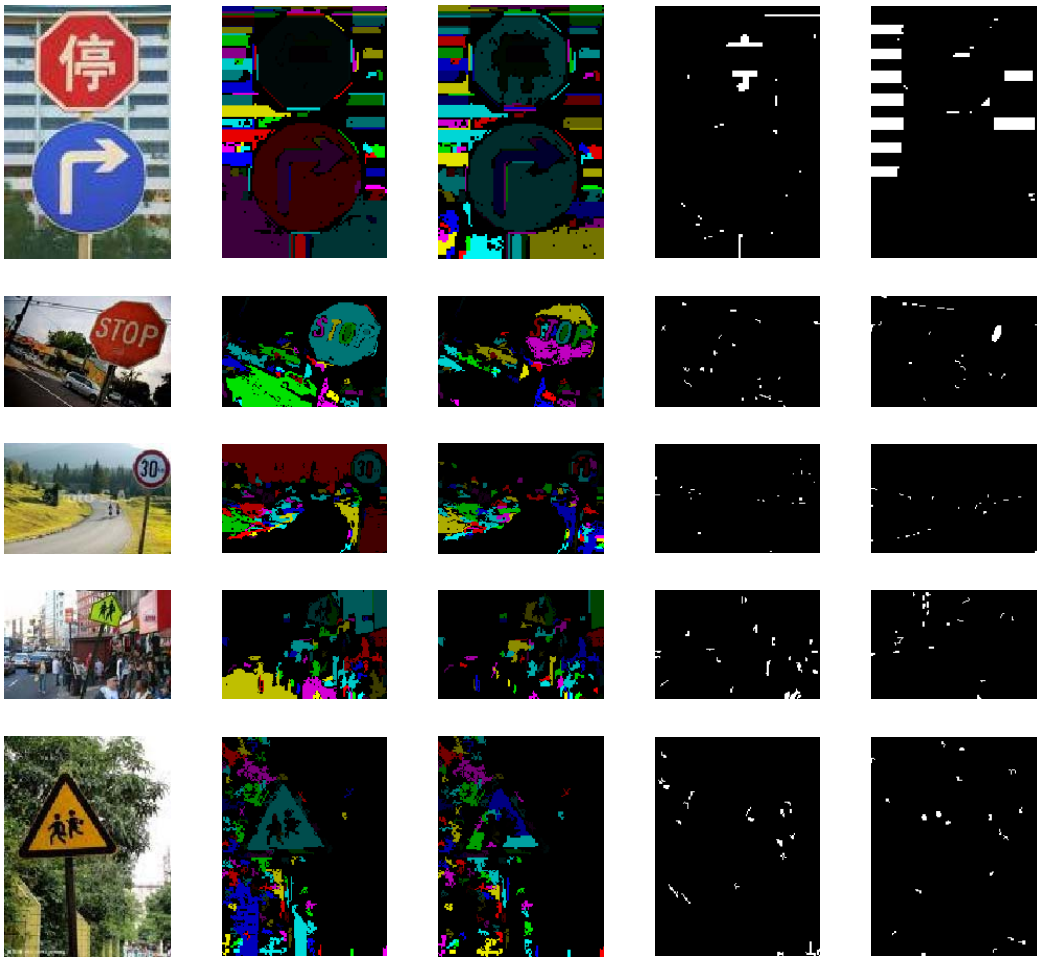
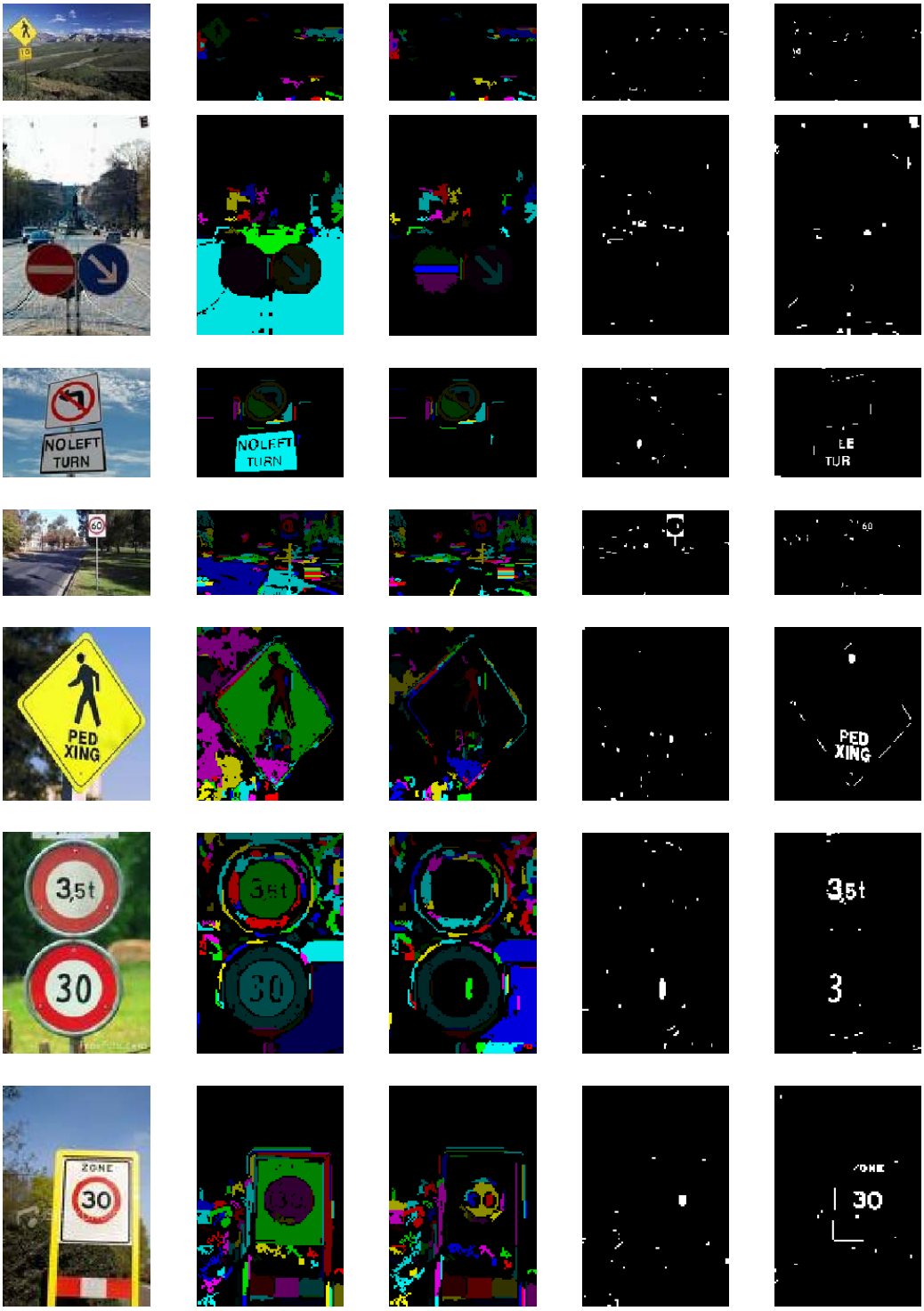


Figure 1: Flowchart of the algorithm

Figure 2: An example of gradually appearing salient regions with threshold increasing





(a): Image (b): MSSR (c): MSSR- (d): MSER+ (e): MSER-

Figure 3: Region Detection Results (MSSR- means MSSR result with internal edge based filtering)