UDC 004.93:12

A. KASHTALIAN
Khmelnytsky National University

SEGMENTATION OF TEXT INFORMATION IN NATURAL SCENE IMAGES

Text detection and recognition is one of the difficult task in the computer vision, in particular in the case of images with a complicated background. The article is devoted to investigation of the task of text recognition in images with non-uniform background, in particular segmentation stage. The segmentation on lines, words and symbols are examined. The lines segmentation approach is based on determination of general intensity or color channels intensity and assumes definition of average intensity of the whole image, looking through every pixel line and definition its intensity, comparison line intensity with average intensity of image, finding of border between text and line spacing by intensity difference. The words segmentation approach is also based on determination of general intensity or intensity of color channels and consist of definition of text line general intensity, looking through every pixel column and definition its intensity, comparison with average text line intensity, finding the border between a word and space by intensity difference. The character segmentation based on finding maximally stable extremal regions (MSER) is suggested. The maximally stable external regions (MSER) feature detector works well for finding text regions because of the stable intensity profiles. This is a method for blob detection in images. The algorithm extracts from an image a number of co-variant regions: a region is a stable connected component of some gray-level sets of the image. The MSER extraction implements the following steps: sweep threshold of intensity from black to white, perform a simple luminance thresholding of the image; extract connected components (extremal regions); find a threshold when an extremal region is "maximally stable", i.e. local minimum of the relative growth of its square. The MSER detector marks out most of the text, it also detects many other stable regions in the images that are not text. These candidates are then filtered using regions geometric properties and stroke width information to exclude non-text objects. The segmentation process allows to extract characters for father classification and to create a dataset necessary for classifier training. The coding has implemented in python and qualitative analysis is performed.

Key words: optical character recognition, maximally stable extremal regions, symbol allocation, intensity

А.С. КАШТАЛЬЯН
Хмельницький національний університет

СЕГМЕНТАЦІЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ НА ЗОБРАЖЕННЯХ З ФОНОМ

Детектування та розпізнавання тексту є одним з складних завдань машинного навчання, зокрема це стосується зображень зі складним фоном. Стаття присвячена дослідженню питання розпізнавання тексту на зображення з неоднорідним фоном, зокрема етапу сегментації. Розглянуто сегментацію текстової інформації на рядки, слова та символи. Спосіб сегментації рядків ґрунтується на визначенні сумарної інтенсивності або інтенсивності каналів всього зображення та передбачає знаходження середньої інтенсивності всього зображення, проходження всіх піксельних рядків зображення та визначення їх інтенсивності, порівняння з сумарною інтенсивністю зображення, знаходження границі розділу рядку тексту та міжрядкового інтервалу за різницею інтенсивності. Спосіб сегментації слів також ґрунтується на визначенні сумарної інтенсивності або інтенсивності каналів і включає знаходження середньої інтенсивності текстового рядка, проходження всіх піксельних стовбців в рядку та визначення їх інтенсивності, порівняння з сумарною інтенсивністю рядка, знаходження границі розділу слова та проміжку між словами за різницею інтенсивності. Запропоновано сегментацію символів на основі знаходження максимально стабільних екстремальних регіонів (MSER). MSER детектор добре працює для знаходження текстових символів в умовах стабільності інтенсивності символьних зображень. Для виділення MSER областей виконуються наступні кроки: проходження порогу інтенсивності від чорного до білого; виконання простої порогової обробки інтенсивності зображення; виділення зв'язаних компонент (екстремальних регіонів); знаходження порогу, за якого екстремальний регіон буде максимально стабільним. MSER детектор достатньо добре позначає області тексту, але він також детектує багато інших стабільних областей на зображенні, які не є текстом. Регіони-кандидати потім фільтруються, використовуючи геометричні властивості та інформацію про ширину лінії для виключення нетекстових областей. Процес сегментації дозволяє виділити символи для подальшої їх класифікації, а також сформувати датасет, необхідний для навчання класифікатора. Експериментальні дослідження виконано з допомогою мови програмування python.

Ключові слова: оптичне розпізнавання тексту, максимально стабільні екстремальні регіони, виділення символів, інтенсивність.

Introduction

The amount of processed information constantly increases. Optical character recognition (OCR) is an important part of artificial intelligence application methods. Its applications are extremely wide; it is video processing, security activity, multimedia libraries, document flow, cartography, etc. An artificial imposed text or a natural text on video, images, documents are a source of important information. Nowadays the recognition of text can be used in even bigger number of application, including people navigation improving, supporting people with feeble sight, machine translation, because of the permanently increasing number of cameras, like smartphone cameras. The task of the recognition of text fragments on contrast images with uniform background is well investigated and successfully solved by now. However, the extraction of text information becomes a challenge in the presence of more complex background structure. Different factors begin to influence result, including text color, space text orientation, existence of similar to text elements.

The increase in indicators of accuracy of text allocation and segmentation in situations of difficult graphic scenes, which are characterized by a non-uniform background, a lack of precise criteria of difference between a background and the text, high probability of various distortions, is crucial for modern applications. The development of modern intellectual technologies of search of the text on video and images has to allow strengthening extent of

implementation of these technologies in different fields of activity farther.

Related Work

Major of works on text segmentation are focused on documents or on constrained contexts, like license plate detection and recognition. However, OCR area is in continuous development and it is possible to mark out several main directions.

One group consists of methods, which use contour information, because each symbol has clear expressed contour structure. Such approaches as skeletonization [1], edge detection and corner detection [2], invariant methods [3] and similar ones are used for text localization in this case. Fast data processing received at a preprocessing stage can represent an uncommon task in a case of images with a difficult background.

Another group of methods are based on color information processing, target areas usually have homogeneous color/intensity and satisfy restrictions on a form and size. The known approaches are histogram method [4], connected component analysis [5] and different algorithms of adaptive binarization including algorithms of Niblack, Sauvola, Chistian, Bernsan, Otsu [2]. Methods allow working with optional font size and optional text direction, but they do not work well enough on images with difficult background, noise and unsharpness, use the big number of heuristics.

One more group of methods is methods, which use textural information, text zones may significantly differ from a background, it allows to use various frequency filters for "images pyramid". The pattern recognition classical methods can be used for definition of target areas, among them support vector machines, neural networks, expert systems etc. [6]. In addition, special methods are used, for example the method of spectrographic structures [7]. These approaches allow processing images with complicated background; have high computing complexity because of necessarily of image scaling. The issue of textural features allocation is described in fundamental works like [8].

The known commercial decisions nevertheless do not give the possibility to receive necessary quality of speed of data processing. Existent methods are combined by the processing scheme: image quality improvement, segmentation for text areas identification, clustering and actually recognition. The algorithms presented in the open sources with use of neural networks and adaptive binarization provide high percent of text areas extraction. The lack of approach with use of neural networks is the high computing complexity, the necessarily of work with large sizes of the training sets and ambiguity at the choice of a neural network architecture. The methods, which use the shaped filters, show the same accuracy due to application of a support vectors approach for data processing.

Lines and Word Segmentation

Text recognition include next important parts: detect text areas in images, segment text areas to symbols, recognize symbols, gather words and sentences. Segmentation part may contain lines segmentation, words segmentation, symbols segmentation. These processes may be accurately separated or combined.

Lines segmentation. The task of line allocation comes down to finding of upper and lower lines sides of a text. Line segmentation is based on the fact that average brightness of lines spacing differs significantly from average brightness of text lines.

The average value of brightness of all pixel lines is

$$s_j = s_j(B) = \frac{1}{n}\sum_{i=1}^{n} b_{ij},$$

n – the number of pixels in one line, b – brightness of a pixel.

The average value of brightness of whole image is

$$s(B) = \frac{1}{m}\sum_{j=1}^{m} s_j(B),$$

m – the number of line in a image.

The brightness of an upper border of a text line can be expressed through average brightness of a whole image $s^t = k^t \cdot s(B)$, where $0 < k^t < 1$ - a coefficient.

The process of lines segmentation consists in consistent looking through a massive of average values $(s_1,\ldots,s_m)$ and identification of a set of indexes pairs $\left(s_i^t, s_i^b\right)$ of pixel lines corresponding upper s_i^t and lower s_i^b borders of the line with a number i. These pairs should satisfy conditions:

1. The condition of upper border of a text line. The beginning of a text line is defined if next conditions are satisfied: a brightness of a current pixel line differs from the border s^t; a brightness of two previous lines differs from this border; a brightness of next three lines differs from a border s^b.

2. The condition of lower border of a text line. The end of area of a steady change of brightness is defined if next conditions are satisfied: the beginning of area was fixed; a brightness of current pixel line differs from a border s^t; a brightness of next pixel line differs from a border s^b; or the beginning of area was fixed; a brightness of three next lines differs from a border s^b.

The set of pairs of upper and lower borders is computed in the described way. The difference between upper and lower indexes is a high of text lines. The found borders cut off overstepped symbols. It is necessary to expand the values of borders. The size of expansion depends on line spacing.

Words segmentation. The input for words segmentation is an image of one text line. This image can be obtained after lines segmentation in the case of multiline text or initial text. Firstly, it needs to perform two transformation of input image:

1. The threshold filter for contrast increasing

$$b_{ij} = \begin{cases} b_{\max}; & b_{ij} > b_0 \\ 0; & b_{ij} < b_0 \end{cases}$$

$i = 1 \ldots n$; $j = 1 \ldots m$; b_0 – brightness threshold.

2. Low-pass filtering.

The words segmentation algorithm is based on a fact that an average brightness in inter-word spaces significantly differs from a brightness inside words.

The average brightness of all pixel columns of initial line image is

$$c_i = c_i(B) = \frac{1}{m} \cdot \sum_{j=1}^{m} b_{ij},$$

m – the high of current line in pixels.

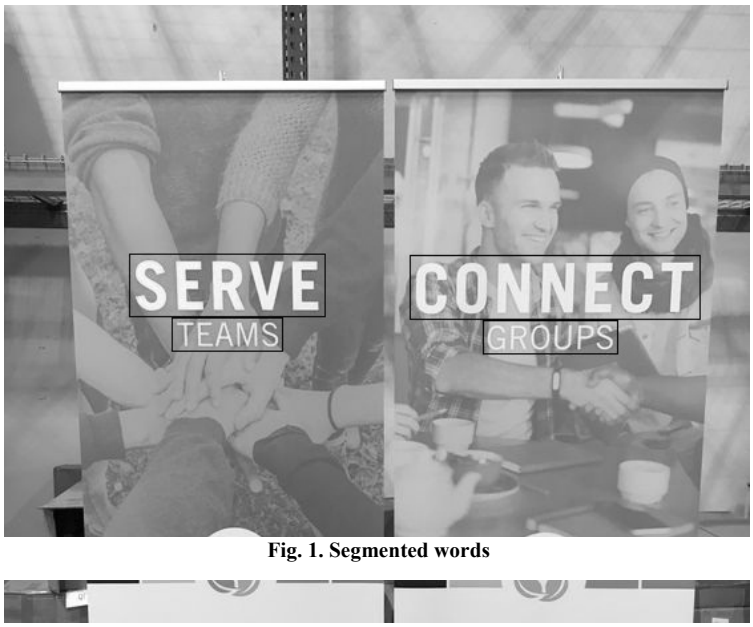


Fig. 1. Segmented words

The average brightness of the line image is

$$c(B) = \frac{1}{n} \cdot \sum_{i=1}^{n} c_i(B),$$

n – the width of a current line in pixels.

The left border of a word (the beginning of a word) is expressed through an average brightness of a line image - $c^l = k^l * c(B)$, where $0 < k^l < 1$ - a coefficient.

The right border of a word (the end of a word) is also expressed through an average brightness of a line image - $c^r = k^r * c(B)$, $0 < k^r < 1$.

The algorithm assumes consistent looking through the set of average brightness values of pixel columns $(c_1 \ldots c_n)$ and finding the set of index pairs (c_i^l, c_i^r) of pixel lines corresponding a left c_i^l and a right c_i^r borders of word *i*. It should satisfy next conditions:

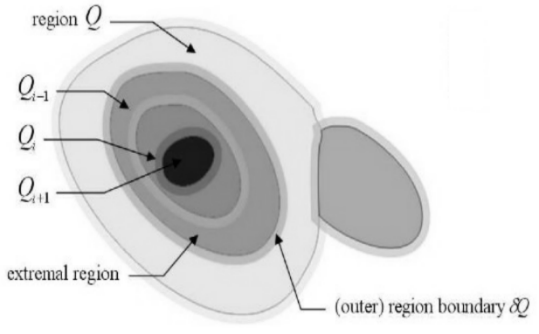
1. The conditions of a left bounder (the beginning of a word). The beginning of a word is considered when: a brightness of current and next pixel columns differ from a left border for a word c^l; a brightness of a previous pixel column differs from this border.
2. The conditions of a right bounder (the end of a word). The end of a word is defined if: the beginning of a word was fixed; a brightness of current and four next pixel columns is different from a brightness border c^r of an inter-word space; a brightness of two previous pixel columns differs from this border.

The example of result of lines and words segmentation is presented in the fig. 1.

Symbol Segmentation with Maximally Stable Extremal Regions

Image *I* is a mapping *I*: $D \subset Z^2 \to S$. Extremal regions are well defined on images if:

1. *S* is totally ordered, i.e. reflexive, antisymmetric and transitive binary relation $\leq$ exists. In this paper only $S = \{0,1,\ldots,255\}$ is considered, but external regions can be defined on e.g. real-valued images (*S*=*R*).
2. An adjacency (neighborhood) relation $A \subset D \times D$ defines. If 4-neighborhoods are used than $p, q \in D$ are adjacent (pAq) if $\sum_{i=1}^{d} |p_i - q_i| \leq 1$.



Region Q is a contiguous subset of *D*, i.e. for each $p, q \in Q$ there is a sequence

$p, a_1, a_2, \ldots, a_n, q$ and $pAa_1, a_i Aa_{i+1}, \ldots, a_n Aq$.

Region (outer) border $\partial Q = \{q \in D \setminus Q : \exists p \in Q : qAp\}$, i.e. the border ∂Q of Q is the set of pixels being adjacent to at least one pixel of Q but not belonging to Q (fig. 2).

Extremely region $Q \subset D$ is a region such that for every $p \in Q, q \in \partial Q$: $I(p) > I(q)$ (maximum intensity region) or $I(p) < I(q)$ (minimum intensity region) [8].

Fig. 2. Maximally Stable Extremal Regions

The maximally stable external regions (MSER) feature detector works well for finding text regions because of the stable intensity profiles. The algorithm of affine invariant intensity extremes implements in the following sequence: start from a local intensity extreme point; go in every direction until the point of extreme of some function *f*, the curve connection the points is the region boundary; compute geometric moments of orders up to 2 for this region; replace the region with a rectangle.

MSER is a method for blob detection in images. The MSER algorithm extracts from an image a number of co-variant regions: an MSER is a stable connected component of some gray-level sets of the image. MSER is based on the idea of taking regions that stay nearly the same through a wide range of thresholds:

- all the pixels below a given threshold are white and all those above or equal are black;
- if we are shown a sequence of threshold images I_t with frame corresponding to threshold *t*, we would see first a black image, then white spots corresponding to local intensity minima will appear then grow larger;
- these white spots will eventually merge, until the whole images is white;
- the set of all connected components in the sequence is the set of all extremely regions.

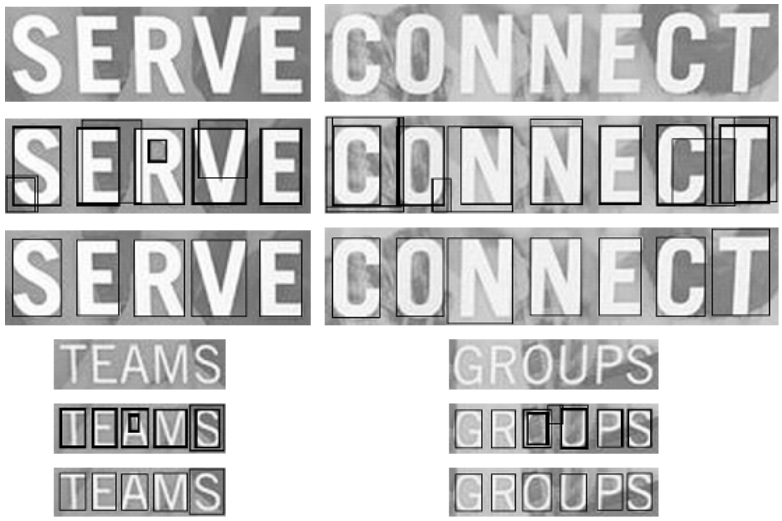


Fig. 3. Symbol segmentation by MSER detector and selection of the best regions-candidates

Optionally, elliptic frames are attached to the MSERs by fitting ellipses to the regions. Those regions descriptors are kept as features.

The word extremal refers to the property that all pixels inside the MSERs have either higher (bright external regions) or lower (dark external regions) intensity than all the pixels on its outer boundary. This operation can be performed by first sorting all pixels by gray value and then incrementally adding pixels to each connected component as the threshold is changed. The area is monitored. Regions such that their variation with the threshold is minimal are defined maximally stable.

The MSER extraction implements the following steps: sweep threshold of intensity from black to white, perform a simple luminance thresholding of the image; extract connected components (extremal regions); find a threshold when an extremal region is "maximally stable", i.e. local minimum of the relative growth of its square. Due to the discrete nature of the image, the region below/above may be coincident with the actual region, in which case the region is still deemed maximal; approximate a region with an ellipse (optional); keep those regions descriptors as features.

Fig. 4. Example of training sample symbols for machine learning classifier

In spite of the fact that MSER detector marks out most of the text, it also detects many other stable regions in the image that are not text. Firstly, geometric properties of text use to filter out non-text regions using simple rules. Several geometric properties are good for discriminating between text and non-text regions, including aspect ratio, eccentricity, Euler number, extent, solidity. It is also necessary to consider

separately the regions located inside other regions and remove repeating ones (fig. 3). In addition, machine-learning approach is used to train text vs non-text classifier as support vectors machine, neural networks etc. It is necessary to create the training set with text and non-text symbols for training a machine learning classifier (fig. 4). A combination of the two approaches produces better results. The next task is the classification of detected symbols. Experiments were conducted with python language.

Conclusion

In this work, a general methodology for text lines, words and characters segmentation in natural scene images is presented. Compare to others this approach is general and can be applied to wide range of images independently from background non-uniformity. The properties of difference intensity and MSER detection were be combined. Intensity difference was used for lines and words segmentation, because they generally have similar features. MSER blob detection was used for characters' detection. The combination of these methods give high quantity of text detection in images with background. The exact number of detection accuracy depends on the quality of input images. The number of detected symbols can be increased due to image preprocessing.

References

1. Pogodin S.V. Allocation and the analysis of skeletons of objects in color pictures / S.V. Pogodin // Software products and systems. – Russia, Tver, 2009. – Volume 2. – P. 42–45.

2. Shapiro L. Computer vision / Shapiro L., Stokman J. – Russia, Moscow, Binom, 2006 – 752 p.

3. Aviles C.C. A robust font recognition using invariant moments / C.C. Aviles, C.J. Villegas, H.J. Ocampo // Proceedings of the 5th WSEAS International Conference on Applied Computer Science – China, Hangzhou, 2006. – P. 114–117.

4. Vinogradov A.N. Extraction and recognition of local objects in space pictures / A.N. Vinogradov, F.V. Kalugin, M.D. Nedev // Aerospace instrument making. – Russia, Moscow, 2007. – Volume 9. – P. 39–45.

5. Feby Ashraf. Connected component clustering based text detection with structure based partition and grouping / Feby Ashraf, V. A. Nurjahan // Journal of Computer Engineering, 2014. – Volume 16, Issue 5. – P. 50–56.

6. Talalaev A.A. Allocation and clustering of text and graphic elements in semitone pictures / A.A. Talalaev, I.P. Tyshchenko, M.V. Hachumov // Artificial intelligence and decision making. – Russia, Moscow, 2008. – Volume 3. – P. 72–84.

7. Fralenko V.P. The analysis of spectrographic textures of Earth remote sensing data / V.P. Fralenko // Artificial intelligence and decision-making. – Russia, Moscow, 2010. – Volume 2. – P. 11–15.

8. Matas J. Robust wide baseline stereo from maximally stable extremal regions / J. Matas, O. Chum, M. Urban, and T. Pajdla // Proc. of British Machine Vision Conference, 2002. – P. 384–396.

Рецензія/Peer review : 13.11.2017 р. Надрукована/Printed :05.12.2017 р.
Рецензент: д.т.н., проф. Полікаровських О.І.