Text Localization in Real-world Images using Efficiently Pruned Exhaustive Search

Lukáš Neumann
Centre for Machine Perception, Dept. of Cybernetics
Czech Technical University, Prague, Czech Republic
neumalu1@cmp.felk.cvut.cz

Jiří Matas
Centre for Machine Perception, Dept. of Cybernetics
Czech Technical University, Prague, Czech Republic
matas@cmp.felk.cvut.cz

***Abstract*—An efficient method for text localization and recognition in real-world images is proposed. Thanks to effective pruning, it is able to exhaustively search the space of all character sequences in real time (200ms on a 640×480 image). The method exploits higher-order properties of text such as word text lines. We demonstrate that the grouping stage plays a key role in the text localization performance and that a robust and precise grouping stage is able to compensate errors of the character detector.**

The method includes a novel selector of Maximally Stable Extremal Regions (MSER) which exploits region topology. Experimental validation shows that 95.7% characters in the ICDAR dataset are detected using the novel selector of MSERs with a low sensitivity threshold.

The proposed method was evaluated on the standard ICDAR 2003 dataset where it achieved state-of-the-art results in both text localization and recognition.

***Keywords*-text localization;real-world images;text-in-the-wild**

I. Introduction

Text localization and recognition in real-world images is an open problem, unlike printed document recognition where state-of-the-art systems are able to recognize correctly more than 99% of characters [1]. Applications of text localization and recognition in real-world images range from automatic annotation of image databases based on their textual content (e.g. Flickr or Google Images), assisting the visually impaired to reading labels on businesses in map applications (e.g. Google Street View).

Existing methods for text localization can be categorized into two different groups - methods based on a sliding window and methods based on regions (characters) grouping. Methods in the first category [2], [3] use a window which is moved over the image and the presence of text is estimated on the basis of local image features. While these methods are generally more robust to noise in the image, their computation complexity is high because of the need to search with many rectangles of different sizes, aspect ratios and potentially rotations. Additionally, support for slanted or perspectively distorted text is limited.

The majority of recently published methods for text localization falls into the latter category [4], [5], [6], [7]. The methods differ in their approach to individual character detection, which could be based on edge detection, character energy calculation or extremal region detection. While the methods are paying great attention to individual character detection, grouping of individual characters into words is performed based on heuristics or graph optimization methods and only unary and pairwise constraints are used.

In this paper, a general and efficient method for text localization and recognition is presented, which thanks to effective pruning is able to group character regions by an exhaustive enumeration of all character sequences. The method exploits higher-order properties of text, which cannot be incorporated into standard graph (or hypergraph) optimization methods where only unary or binary features are used. We demonstrate that the grouping stage plays a key role in the text localization performance and that even a character detector with a lower precision is sufficient if the grouping stage is accurate.

As a second contribution, an extended version of Maximally Stable Extremal Regions (MSERs) [8] called MSER++ is introduced. Experimental evaluation shows that 95.7% characters are detected as MSER++, which is a significant improvement over standard MSER (84.0%) as used in our previous method [6].

The rest of the document is structured as follows: In Section II, the proposed method is described, in Section III experimental evaluation is performed and the paper is concluded in Section IV.

II. Text localization

A. Character grouping search space

Let $\mathbf{I}$ denote an image of n pixels and let $P(\mathbf{I})$ denote set of all subregions of the image $\mathbf{I}$. Let s^L denote an arbitrary sequence of non-repeating image subregions $s^L = (r_1, r_2, \ldots, r_L) : r_i \in P(\mathbf{I}),\ r_i \neq r_j\ \forall i, j$ of length L, let $\mathcal{S}^L = \bigcup_{i=1}^{n} s^i$ denote set of all sequences of length L and let $\mathcal{S}$ denote set of all sequences of lengths up to n $\mathcal{S} = \bigcup_{i=1}^{n} \mathcal{S}^i$. Given a verification function $v : \mathcal{S} \rightarrow \{0, 1\}$, the set of estimates (*words*) $\mathcal{E}$ is defined as $\mathcal{E} = \{w \in \mathcal{S} : v(w) = 1\}$. The methods for text localization aim to find an optimal verification function $v^*(s)$ so that f-measure of precision $p = \frac{|\mathcal{E} \cap \mathcal{T}|}{|\mathcal{E}|}$ and recall $r = \frac{|\mathcal{E} \cap \mathcal{T}|}{|\mathcal{T}|}$ is maximized, where $\mathcal{T}$ denotes set of words in the ground truth.

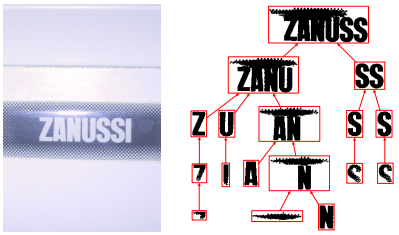


Figure 1. MSER lattice induced by the inclusion relation. Only certain nodes correspond to characters

Table I
INDIVIDUAL CHARACTER DETECTION RATE USING DIFFERENT VARIANTS OF MSER

MSER	root only (%)	complete tree (%)
Greyscale	74.9	93.1
Red channel	72.8	93.2
Green channel	74.9	93.5
Blue channel	72.5	87.1
Combined	84.0	**95.7**

B. Extended Maximally Stable Extremal Regions

The cardinality of the set $\mathcal{S}$ is exponential in the number of pixels in the image, thus it is infeasible to exhaustively search the whole set in order to obtain an optimal solution even if we assume that an optimal verification function $v^*(s)$ exists and can be efficiently calculated. Assuming that each character is a contiguous region of the image $\mathbf{I}$ (which implies that dots and accents have to be handled separately), the set $\mathcal{S}$ can be limited to a set of sequences of contiguous regions without any loss in performance.

Zimmerman and Matas [9] showed that individual characters are often Category Specific Extremal Regions (CSERs) and Donoser et al. [10] further demonstrated that characters are detected as Maximally Stable Extremal Regions (MSERs) [8]. In [6], we show that detection rate of MSERs is improved if multiple projections are used.

In this paper, we extend this approach by using whole tree of MSER lattice induced by the inclusion relation, in contrast to [6] where only root nodes (i.e. supremums of the MSER lattice) were considered which implied that a high MSER margin had to be used to maximize the number of root nodes which correspond to letters. If a lower margin is used, the MSER detector finds more regions but only certain regions correspond to characters. As shown in Figure 1, smaller MSERs are embedded into bigger ones, thus forming a tree where only certain combinations of nodes can be selected as letters, because in a word one letter cannot be embedded into another. We refer to individual nodes of the MSER tree as MSER++ to emphasize that multiple projections (gray, red, green and blue channel) are used and the internal tree structure is taken into account.

C. Exhaustive search

Let $\mathcal{M}$ denote the set of MSER++ in the image $\mathbf{I}$. Even though the cardinality of $\mathcal{M}$ is linear in number of pixels, the cardinality of the set $\mathcal{S}$ of all sequences is still exponential (the complexity has decreased from 2^{2^n} to 2^n only).

Let $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_n$ denote "upper-bound" verification functions which determine whether s^L is a subsequence of a text sequence or a text sequence itself

$$\hat{v}_L(s^L) = 1 \iff \exists s' : s^L \subseteq s',\ v(s') = 1 \qquad (1)$$

It follows that the enumeration of $\mathcal{E} = \{w \in \mathcal{S} : v(w) = 1\}$ can be equivalently defined as finding the set of unextendable sequences

$$\hat{\mathcal{E}} = \bigcup_{L=1}^{n} \{w \in \mathcal{E}^L :\ \forall s' \supset w \in \mathcal{E}^{L+1}\ \hat{v}_{L+1}(s') = 0\} \qquad (2)$$

where $\mathcal{E}^1, \mathcal{E}^2, \ldots, \mathcal{E}^n$ denote sets of text (sub)sequences of length L

$$\begin{aligned}
\mathcal{E}^1 &= \left\{r \in \mathcal{M} \mid \hat{v}_1(r) = 1\right\} \\
\mathcal{E}^2 &= \left\{(r_1, r_2) \mid r_1, r_2 \in \mathcal{E}^1, r_1 \neq r_2, \hat{v}_2(r_1, r_2) = 1\right\} \\
\mathcal{E}^3 &= \left\{(r_1, r_2, r_3) \mid (r_1, r_2), (r_2, r_3) \in \mathcal{E}^2,\right. \\
&\quad \left. r_i \neq r_j\ \forall i, j, \hat{v}_3(r_1, r_2, r_3) = 1\right\} \quad \ldots \\
\mathcal{E}^n &= \left\{(r_1, r_2, \ldots, r_n) \mid (r_1, r_2, \ldots, r_{n-1}),\right. \\
&\quad \left. (r_2, r_3, \ldots, r_n) \in \mathcal{E}^{n-1}, r_i \neq r_j\ \forall i, j, \hat{v}_n(r_1, r_2, \ldots, r_n) = 1\right\}
\end{aligned} \qquad (3)$$

This decomposition allows efficient pruning of the exhaustive search, because non-text subsequences are excluded without a need to build a complete sequence, which prevents from a combinatorial explosion of enumerating the $\mathcal{S}^L$ sets of all sequences of length L.

D. Verification functions

The choice of upper-bound verification functions $\hat{v}_L$ is crucial for the proposed method. Since the optimal verification function $v^*(s)$ is not known, the upper-bound verification functions $\hat{v}_L$ have to be approximated. The key criteria for the approximation is achieving high recall while rejecting as many non-text sequences as possible to prune the search and limit the size of the $\mathcal{E}^L$ sets.

The function $\hat{v}_1(r)$ is a SVM character classifier, which determines whether the region is a character or not based on a set of region measurements (height ratio, compactness, etc.) - see [6]. The function is scale invariant, but not rotation invariant so possible rotations had to be included in the training set. On average, the $\hat{v}_1$ function correctly includes 83% of text regions whilst it correctly excludes 93% of non-text regions such as plants, trees or other random textures.

The $\hat{v}_2(r_1, r_2)$ function consists of pairwise rules which compare measurements of the two regions. The rules require that height ratio, centroid angle and region distance normalized by region width fall within a given interval obtained in a training stage (similar binary rules have been used in many previous works [5], [6], [7], [4]).

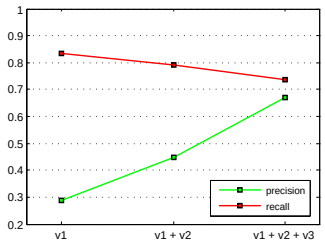


Figure 2. The effects of applying the verification functions $\hat{v}_1$, $\hat{v}_2$ and $\hat{v}_3$ on individual character localization performance

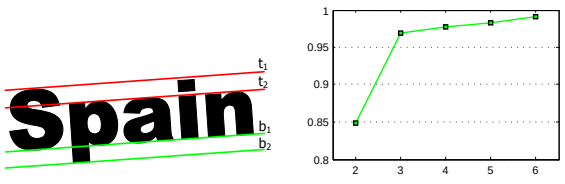
Table II
AVERAGE VERIFICATION FUNCTION CHARACTERISTICS

Function	precision (%)	recall (%)	pruned (%)	total time (s)
$\hat{v}_1$	28.9	83.2	93.3	0.21
$\hat{v}_2$	61.6	94.2	97.0	0.01
$\hat{v}_3$	78.5	87.2	37.2	0.12

In the proposed method, a new rule which exploits the structure of MSER lattice is added. As demonstrated in Figure 1, two regions cannot be in the same sequence if there is a (transitive) parent-child relationship between them, as in this case the first region is embedded into the second one or vice-versa, which is extremely rare for standard text.

In experiments, the $\hat{v}_2$ function pruned out on average 97% of region pairs in an image, but still a significant number of region pairs which are not text passed through (precision is only 62%) - see Table II. This is caused by the fact that the individual measurements on two regions can greatly differ (for instance, the height ratio between the leading capital letter and following lower-case letters can be greater than 3 in some words, the color of two subsequent letters can differ a lot because of lighting conditions, etc.), so a very conservative (i.e. large) interval has to be used to support this variety of texts.

Since the implemented pairwise rules are not sufficiently selective, higher-order features have to be used to reduce the number of false positives. One of such features is based on the observation that letters in a word can be fitted by one or more top and bottom lines (see Figure 3a) and distance of individual letters from these lines is limited (subject to normalization by region height). We refer to this set of lines as *word text lines*.



(a)

(b)

Figure 3. Word text line parameters (a). Dependence of text line parameters estimate accuracy on sequence length (b)

Word text lines estimate $\tau = (t^1, t^2, b^1, b^2, x^{\min}, x^{\max}, h)$ is obtained by inferring a direction k of the text by fitting bottom points using Least-Median of Squares and then fitting top and bottom points of all regions with at most two

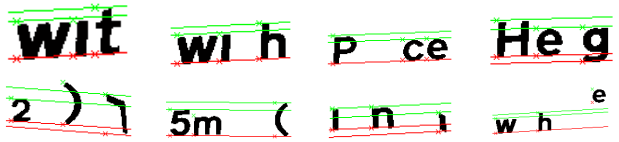


Figure 4. Word text line estimates and triplets accepted (top row) and rejected (bottom row) by the $\hat{v}_3$ function. Top points and lines marked green, bottom points and lines marked red

top (t^1, t^2) and bottom (b^1, b^2) lines with inferred direction k in order to obtain minimal square error (see Figure 4). Variables $x^{\min}$, $x^{\max}$, $h^{\max}$ denote left and right boundary of the word and word height.

In order to obtain a direction of the text, at least 2 regions are needed, but this estimate can be very inaccurate (e.g. fitting bottom points of letters "ly" will result in an incorrect direction, because bottom points of each letter 'sit' on a different bottom line). If three regions are used the estimate is more accurate (see Figure 3b). This fact is exploited by the verification function $\hat{v}_3(r_1, r_2, r_3)$ which creates a word text line estimate τ for given triplet and then verifies that the estimate is valid (mutual vertical distance of the text lines is constrained based on thresholds created during training) and that distance of all three regions from τ is within an interval obtained in a training stage. The recall of after applying the $\hat{v}_3$ function is 87% and 37% of region pairs are pruned out.

The concept of word text lines was used for baseline estimation in printed document analysis [11] and was also applied to text localization in [12]. In the proposed method, only triplets of regions are always used to infer these parameters, in contrast to previous methods where these parameters are estimated on whole words.

As demonstrated in Figure 3b, increasing the number of regions in a sequence does not significantly improve the estimate, which suggests that the geometrical parameters of the word apply to all its subsequences as well. Based on this observation, the verification functions $\hat{v}_4, \ldots, \hat{v}_n$ are approximated by verifying that the text line parameters of all subsequences of length 3 are consistent:

$$\hat{v}_L(s^L) = 1 \iff \forall s_1^3, s_2^3 \subset s^L : d(s_1^3, s_2^3) < \theta \quad (4)$$

The distance $d(s_1^3, s_2^3)$ of two triplets (see Equation 5) is defined as the largest normalized vertical difference of their text line parameter estimates τ at their boundary points. The function $\hat{v}_L(s^L)$ accepts the sequence s if distance between all triplets in the sequence s is below a threshold θ, which is a parameter of the method obtained during training.

Only the smallest distance is taken into account for top lines as some triplets may contain incomplete set of text lines - for instance in the word "Bear" the triplets "Bea", "Bar" and "Ber" have two top lines because of the capital letter "B", whereas the triplet "ear" has only one top line, which can match to any of the two top lines in the remaining triplets. The same argument applies to bottom lines (e.g. "space") and the two situations can even occur at the same

Figure 5. Exhaustive search for text (sub)sequences. Each sequence is displayed by its region centroids (with random noise to avoid overlapping)

time (e.g. "Gray"), however a bottom line is never matched to a top line or vice-versa.

$$d(s_1^3, s_2^3) = d(\tau_1, \tau_2) = \max\left(\min_{i,j=1\ldots2} \delta(t_1^i, t_2^j), \min_{i,j=1\ldots2} \delta(b_1^i, b_2^j)\right) \quad (5)$$

$$\delta(a,b) = \frac{\max\left(|a(x) - b(x)|, |a(x') - b(x')|\right)}{h}$$
$$x = \min(x_1^{\min}, x_2^{\min}) \quad x' = \max(x_1^{\max}, x_2^{\max})$$
$$h = \max(h_1, h_2) \quad (6)$$

In order to overcome low recall of the $\hat{v}_1$ function, the text localization is performed twice: in the first run, all verification functions are taken into account to build initial text line hypotheses. In the second run, v_1 is forced to 1 and only existing text line hypotheses are taken into account (using the hypotheses-verification framework [6]). The recall of $\hat{v}_2, \ldots, \hat{v}_n$ is not as crucial as one region can be present in multiple subsequences.

The verification function approximation does not guarantee that one region is an element of one sequence only. If this situation occurs, the longer sequence is selected and the other conflicting sequences are discarded. This can be seen as a voting process where each sequence votes for its direction and the most significant direction wins. This process effectively eliminates false positives which are not consistent with text line direction (see Figure 5, bottom-right). Let $\mathcal{E}'$ denote a set of estimates $\mathcal{E}$ without conflicting sequences.

In the proposed method, only sequences longer than 3 regions are accepted because of the low individual precision of the $\hat{v}_1$ and $\hat{v}_2$ functions and the inability to utilize the text line geometric features with individual characters or character pairs.

Table III
TEXT LOCALIZATION (TOP) AND RECOGNITION (BOTTOM) RESULTS ON THE ICDAR 2003 DATASET

method	precision	recall	f
proposed method	**0.65**	**0.64**	**0.63**
Hinnerk Becker [14]	0.62	0.67	0.62
Alex Chen [14]	0.60	0.60	0.58
Neumann and Matas [6]	0.59	0.55	0.57
proposed method	**0.72***	**0.62***	**0.67***
Epshtein et al. [5]	0.73*	0.60*	0.66*
Pan et al. [4]	(0.71)	(0.67)	N/A
Zhang et al. [7]	(0.73)	(0.62)	N/A

method	precision	recall	f
proposed method	**0.42**	**0.41**	**0.41**
Neumann and Matas [6]	0.42	0.39	0.40

III. EXPERIMENTS

The proposed method was evaluated on the most cited ICDAR 2003 dataset [13], which contains 249 images with text of varying sizes and positions.

The standard evaluation protocol defined in [13] was used. The protocol uses words as the unit for comparison, where bounding boxes of words output by the evaluated method $\mathcal{E}$ (*estimates*) are compared to the ground truth T (*targets*). The protocol uses the notion of a flexible match of a region r in a set of regions $\mathcal{R}$ as $m(r, \mathcal{R}) = \max_{r' \in \mathcal{R}} m_p(r, r')$, where $m_p(r, r')$ denotes the area of intersection divided by the area of the minimum bounding box containing both rectangles. Precision and recall of text localization are defined as

$$p_l = \frac{\sum_{r_e \in \mathcal{E}} m(r_e, \mathcal{T})}{|\mathcal{E}|} \quad r_l = \frac{\sum_{r_t \in \mathcal{T}} m(r_t, \mathcal{E})}{|\mathcal{T}|} \quad (7)$$

and a standard f-measure is used to combine both figures.

All performance measures are calculated on each image independently and then an average value over all images is taken as performance of the method. The proposed method achieves precision of 0.65 and recall of 0.64, which outperforms the existing methods as shown in Table III.

This performance measure was used in the ICDAR 2003 and 2005 competitions [13], [14], however papers presented later deviate from the original performance measure. In [5], only one precision and recall value over the whole set of estimates and targets is calculated (marked with an asterisk in Table III for comparison), which gives higher weight to images with more words, which typically leads to better results on the ICDAR dataset as the more challenging images in the dataset contain only small number of words. Other papers [7], [4] use whole text lines for evaluation, so direct comparison is not possible (results given in parentheses in Table III).

The localization output of the proposed method was passed to recognition modules of the hypotheses-verification

Figure 6. Text localization and recognition examples on the ICDAR 2003 dataset. The performance measure has the format $(p_l, r_l; p_r, r_r)$. Notice the low number of false positives despite textures in the images and robustness against blur and reflections. Incorrectly recognized letters marked red

Figure 7. Limitations of the proposed method. Reflection of a flash is too strong so the letter "n" is not detected as an MSER (left). An unsupported text line shape and letters written on glass not detected as MSERs (middle). Multiple letters joint into one region (right)

framework [6]. Table III shows text recognition precision (p_r) and recall (r_r), which are only slightly improved over the previous method, because the text recognition evaluation uses very strict metric, so even a significant improvement in text localization does not guarantee that significantly more words will be recognized without any mistake.

IV. Conclusions

An efficient method for text localization in real-word images was introduced. It was demonstrated that suitable selection of verification functions that control the grouping allows exhaustive search of the space of all character sequences to such an extent that the text can be localized and recognized in real time. The method exploits higher-order features, which significantly improves its performance and accuracy.

Additionally, the method includes a novel selector of MSERs which thanks to exploiting region topology allows using lower margin for detection, which improved individual character detection rate from 84.0% to 95.7% without any impact on calculation complexity.

On the highly cited ICDAR dataset [13], the method achieved precision of 0.65 and recall of 0.64 which represents state-of-the-art results in text localization. The precision 0.42 and recall 0.41 of text recognition is also better than our previous method, however the improvement is only marginal as recognition modules are identical to [6]. On a standard PC, the text localization and recognition took on average 830ms per image on the ICDAR dataset (200ms on average for 640×480 images).

Acknowledgment

The authors were supported by EC project FP7-ICT-247022 MASH, by Czech Government reseach program MSM6840770038 and by CTU Grant Agency project SGS10/069/OHK3/1T/13.

References

[1] X. Lin, "Reliable OCR solution for digital content remastering," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Dec. 2001.

[2] X. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," *CVPR*, vol. 2, pp. 366–373, 2004.

[3] R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," *Circuits and Systems for Video Technology*, vol. 12, no. 4, pp. 256 –268, 2002.

[4] Y.-F. Pan, X. Hou, and C.-L. Liu, "Text localization in natural scene images based on conditional random field," in *ICDAR 2009*. IEEE Computer Society, 2009, pp. 6–10.

[5] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR 2010*, pp. 2963 –2970.

[6] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *ACCV 2010*, ser. LNCS 6495, vol. IV, November 2010, pp. 2067–2078.

[7] J. Zhang and R. Kasturi, "Character energy and link energy-based text extraction in scene images," in *ACCV 2010*, ser. LNCS 6495, vol. II, November 2010, pp. 832–844.

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

[9] J. Matas and K. Zimmermann, "A new class of learnable detectors for categorisation," in *Image Analysis*, ser. LNCS, 2005, vol. 3540, pp. 541–550.

[10] M. Donoser, H. Bischof, and S. Wagner, "Using web search engines to improve text recognition," in *ICPR 2008*, pp. 1 –4.

[11] T. Caesar, J. Gloger, and E. Mandler, "Estimating the baseline for written material," in *ICDAR 1995*, vol. 1, pp. 382–385.

[12] L. Neumann and J. Matas, "Estimating hidden parameters for text localization and recognition," in *Proc. of 16th CVWW*, February 2011, pp. 29–26.

[13] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *ICDAR 2003*, 2003, p. 682.

[14] S. M. Lucas, "Text locating competition results," *Document Analysis and Recognition, International Conference on*, vol. 0, pp. 80–85, 2005.