16^{th} *Computer Vision Winter Workshop*
Andreas Wendel, Sabine Sternig, Martin Godec (eds.)
Mitterberg, Austria, February 2-4, 2011

Estimating hidden parameters for text localization and recognition

Lukáš Neumann
neumalu1@cmp.felk.cvut.cz

Jiří Matas
matas@cmp.felk.cvut.cz

Center for Machine Perception
Department of Cybernetics
Czech Technical University
Prague, Czech Republic

Abstract. *A new method for text line formation for text localization and recognition is proposed. The method exhaustively enumerates short sequences of character regions in order to infer values of hidden text line parameters (such as text direction) and applies the parameters to efficiently limit the search space for longer sequences. The exhaustive enumeration of short sequences is achieved by finding all character region triplets that fulfill constraints of textual content, which keeps the proposed method efficient yet still capable to perform a robust estimation of the hidden parameters in order to correctly initialize the search. The method is applied to character regions which are detected as Maximally Stable Extremal Regions (MSERs).*

The performance of the method is evaluated on the standard ICDAR 2003 dataset, where the method outperforms (precision 0.60*, recall* 0.60*) a previously published method for text line formation of MSERs.*

1. Introduction

Text localization and recognition in images of real-world scenes is still an open problem, which has been receiving significant attention in the last decade [12, 1, 5, 4, 10, 3]. In contrast to text recognition in documents, which is satisfactorily addressed by state-of-the-art OCR systems [6], no efficient method for scene text localization and recognition has been yet published.

Methods for text localization are based on two approaches: sliding windows and connected component analysis. The methods based on sliding windows [2] are more robust to noise, but they have high computational complexity (scanning whole image with windows of multiple sizes is required) and they cannot detect slanted or perspectively distorted text. That is why methods based on individual region detection and subsequent connected component analysis are getting more attention in the text localization community [5, 4, 10]. On the most cited dataset (ICDAR 2003 [8]) the methods based on connected component analysis achieve state-of-the-art results in text localization [11].

In this paper, we present a text line formation method, which groups Maximally Stable Extremal Regions (MSERs) [9] representing characters into text lines. The main contribution of this work is an ability to exhaustively enumerate short sequences of character regions in order to infer values of hidden text line parameters (such as text direction) and subsequently applying the parameters to efficiently limit the search space for longer sequences. The exhaustive enumeration of short sequences is achieved by finding all character region triplets that fulfill constraints of textual content, which keeps the proposed method efficient yet still capable to perform a robust estimation of the hidden parameters in order to correctly initialize the search. The method was evaluated using the hypotheses-verification framework for text localization and recognition published by Neumann and Matas [10], where the heuristic text line formation stage was replaced by the proposed method.

The rest of the document is structured as follows: In Section 2, hidden text line parameters used by the proposed method are defined. Section 3 describes the proposed method for text line formation. Performance evaluation of the proposed method is pre-

sented in Section 4. The paper is concluded in Section 5.

2. Hidden text line parameters

It can be observed that text in real-world images follows a certain structure. The structure is not as strict as in the case of text in printed documents, but it is possible to make certain observations at least on the level of individual words; text parameters such as character height, character color, spacing between individual characters have only limited number of distinct values inside a single word. Moreover each word (and possibly more than one word) has an implied direction in which all characters are laid out.

In this paper, we refer to all such parameters as *hidden text line parameters* (or just *hidden parameters*). The initial values of the hidden parameters are obtained by exhaustively enumerating all region triplets and then the inferred values are used to limit the search space during next steps of the text formation. The hidden text line parameters used by the proposed method are height ratio (Section 2.1), centroid angle (Section 2.2) and text direction (Section 2.3).

2.1. Height ratio

The height of two following letters in a word is constrained to a limited interval. In order to express this relation, the height ratio hr between two characters c^1 and c^2 is introduced as

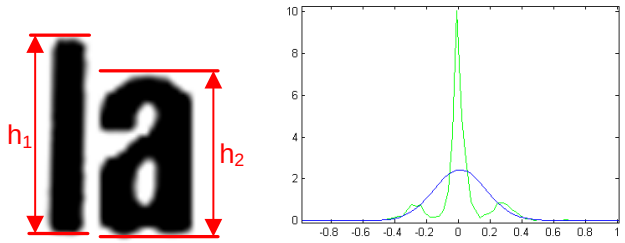
$$\mathsf{hr}(c^1, c^2) = \log \frac{h_1}{h_2} = \log \frac{c_b^1 - c_t^1}{c_b^2 - c_t^2} \qquad (1)$$

where c_t^i and c_b^i denote top and bottom co-ordinate of a bounding box of the character c (see Figure 1a). The measurement is scale invariant, but it is not rotation invariant, which implies that various rotations had to be included in the training set.

Figure 1b depicts the normalized histogram of height ratio values in the training set and their inferred approximation using a Gaussian Mixture Model.

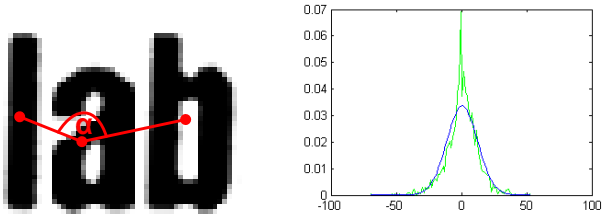
2.2. Centroid angle

Given a sequence of three following letters in a word, the angle between lines connecting their centroids (see Figure 2a) is also constrained to a limited interval. The centroid angle ca of three characters c^1, c^2 and c^3 is defined as



(a) (b)

Figure 1. Height ratio. (a) Measurement example. (b) Normalized histogram (green) and inferred Gaussian Mixture Model M_{hr} (blue)



(a) (b)

Figure 2. Centroid angle. (a) Measurement example. (b) Normalized histogram (right, green) and inferred Gaussian Mixture Model M_{ca} (blue)

$$\mathsf{ca}(c^1, c^2, c^3) = \left| \arctan\left(\frac{c_{cy}^1 - c_{cy}^2}{c_{cx}^1 - c_{cx}^2}\right) - \arctan\left(\frac{c_{cy}^2 - c_{cy}^3}{c_{cx}^2 - c_{cx}^3}\right) \right| \qquad (2)$$

where c_{cx}^i (c_{cy}^i) denotes horizontal respectively vertical co-ordinate of a centroid of the character c^i. The measurement is both scale and rotation invariant.

Figure 2b depicts the normalized histogram of centroid angle values in the training set and their inferred approximation using a Gaussian Mixture Model.

2.3. Text direction

The structure of text in real-world images exhibits higher-order properties, which cannot be fully captured by measurements which are defined only using pairs or triplets of individual characters (such as the parameters in Sections 2.1 and 2.2).

In this paper we introduce a set of parameters called *text direction* to capture higher-order structure of text, which exploits an observation that the top and bottom boundaries of individual characters in a word can be fitted by a line. Depending on which letters form the word, each word has either 1 or 2 top lines (see Figure 3), depending whether only upper-case or both upper-case and lower-case letters are

present in the word. Let $t_1(x)$ and $t_2(x)$ denote vertical position of first respectively second top line at point x. The same observation applies to the bottom lines where either 1 or 2 lines are present, depending whether underline characters such as "y" or "g" are present or not. Let $b_1(x)$ and $b_2(x)$ again denote vertical position of the bottom lines at point x. *Text direction* T is then defined as quaternion (t_1, t_2, b_1, b_2).

Given a text direction T, *text direction distance* of a character c is defined as

$$d(c,T) = \max\Big(\min(|t_1(c_l) - c_t|, |t_2(c_l) - c_t|), \min(|b_1(c_l) - c_b|, |b_2(c_l) - c_b|)\Big) \quad (3)$$

where c_t, c_l and c_b denote top, left and bottom coordinate of a bounding box of the character c.

Mutual position of the lines is not arbitrary either. An assumption was made that these lines are parallel, because height of individual characters in a single word is assumed to be constant and effects caused by perspective distortion in a single word are marginal. Let $D(a(x), b(x)) = |a(x) - b(x)|$ denote vertical distance between lines a and b at horizontal co-ordinate x. Since it was assumed that the lines are parallel, the distance D does not depend on the horizontal position and we can simply write $D(a, b)$ for distance between lines a and b.

In order to express the constraints for mutual vertical distance of the lines, a height of a top bend h_t, a middle bend h_m and a bottom bend h_b is defined (see Figure 3) as

$$h_t(T) = D(t_1, t_2) \quad (4)$$

$$h_m(T) = D(\max(t_1, t_2), \min(b_1, b_2)) \quad (5)$$

$$h_b(T) = D(b_1, b_2) \quad (6)$$

In order to make the text direction parameters scale invariant, they are normalized using a maximal height of a character in the word $h_{\max}$:

$$\bar{d}(c) = \frac{d(c)}{h_{\max}} \quad (7)$$

$$\bar{h}_t(T) = \frac{h_t(T)}{h_{\max}} \quad (8)$$

$$\bar{h}_m(T) = \frac{h_m(T)}{h_{\max}} \quad (9)$$

$$\bar{h}_b(T) = \frac{h_b(T)}{h_{\max}} \quad (10)$$

As shown in Figure 4 the variance of text direction distance $\bar{d}(c)$ measured on the training set is relatively small, which suggests that this parameter can be used as a feature to distinguish between textual and non-textual structures.

Figure 3. Text direction - top lines (red) and bottom lines (green)

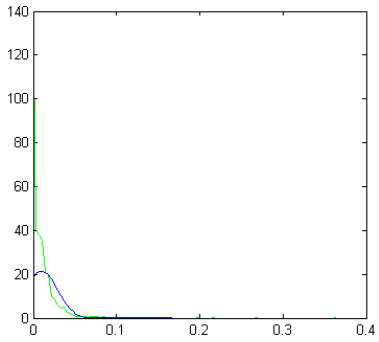


Figure 4. Text direction distance $\bar{d}(c)$ - histogram (green) and inferred Gaussian Mixture Model M_d (blue)

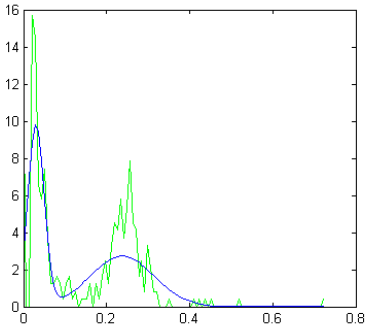


Figure 5. Top band height $\bar{h}_t$ - histogram (green) and inferred Gaussian Mixture Model M_{tb} (blue)

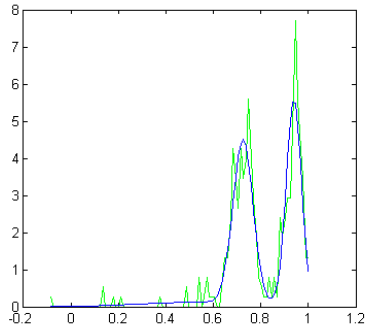


Figure 6. Middle band height $\bar{m}_t$ - histogram (green) and inferred Gaussian Mixture Model M_{mb} (blue)

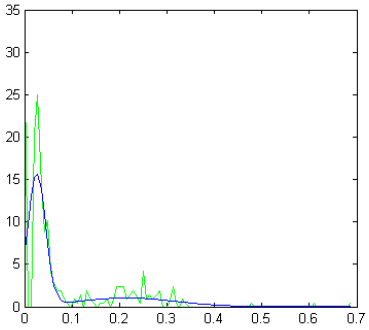


Figure 7. Bottom band height $\bar{b}_t$ - histogram (green) and inferred Gaussian Mixture Model M_{bb} (blue)

```
Procedure la(cc)
 tp := top points of all chars in cc
 bp := bottom points of all chars in cc
 ap := fit bp by a line using Least-Median Squares
 k := tangent of ap

 t1,t2 := fit(tp, k)
 b1,b2 := fit(bp, k)
 T := (t1, t2, b1, b2)
 return T

Procedure fit(points, k)
 bestError := Inf
 for each p,q in points
  line1 := line through p with tangent k
  line2 := line through q with tangent k

  error := 0
  for each r in points
   dist := (min(line1(r[x]),line2(r[x]))-r[y])^2
   error := error + dist

  if error < bestError
   bestError := error
   l1 := line1
   l2 := line2

 return (l1, l2)
```

Figure 8. Pseudo-code of the text direction approximation procedure $\mathsf{la}(cc)$

In order to obtain the text direction T from a sequence of characters $cc = c^1, c^2 \dots c^n$ a procedure $\mathsf{la}(cc)$ is introduced (see Figure 8). The example output of the procedure is shown in Figure 9.

3. Text line formation

3.1. Region graph

Individual characters are obtained by detecting Maximally Stable Extremal Regions (MSERs) [9] and then including only the MSERs which are classified as characters using a trained classifier, as proposed by Neumann and Matas [10].

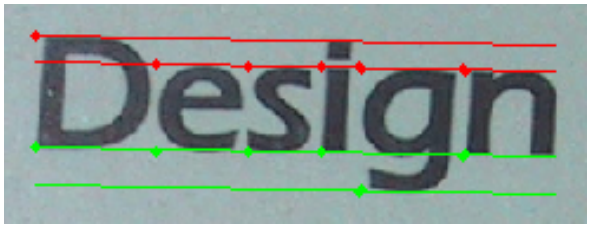


Figure 9. Sequence of characters cc with marked top (red) and bottom (green) points and text direction (top lines - red, bottom lines - green) obtained using the procedure $\mathsf{la}(cc)$

Figure 10. Region graph (initial configuration without any edge labeling)

Let $G = (V, E)$ denote the region graph. The set of vertices V corresponds to the set of character MSERs found in the image. The set of edges E is formed in the following matter: For each vertex, edges to 3 nearest neighboring vertices to the right are created (whilst excluding edges whose centroid angle α is above $40°$). The distance between two vertices is measured as the distance between their centroids. Figure 10 shows an example of such a graph.

3.2. Graph energy

Let $f : E \to \{0, 1\}$ denote a configuration of the region graph G. The text localization task is formulated as finding the best configuration f^* of given graph G such that graph energy $\mathcal{E}(G, f)$ is minimal:

$$f^* = \underset{f}{\operatorname{argmin}}\, \mathcal{E}(G, f) \tag{11}$$

The energy $\mathcal{E}$ is composed of the following weighted components

$$\begin{aligned}\mathcal{E}(G, f) = \alpha_1 \mathcal{E}_{hr}(G, f) + \alpha_2 \mathcal{E}_{ca}(G, f) \\ + \alpha_3 \mathcal{E}_{d}(G, f) + \alpha_4 \mathcal{E}_{la}(G, f)\end{aligned} \tag{12}$$

where $\mathcal{E}_{hr}$ denotes energy of character height ratios (see Section 2.1), $\mathcal{E}_{ca}$ denotes energy of character centroid angles (see Section 2.2) and $\mathcal{E}_d$ ($\mathcal{E}_{la}$) denotes energy of text direction distances and energy of line approximation respectively (see Section 2.3). Coefficients α_i then denote non-negative weights, which in our setup were all set to 1 in order to give each energy an identical weight. The individual energy components are defined using a Gaussian Mixture Model (GMM) approximation, which was created using the training dataset (as shown in Figures 1, 2, 4, 5, 6 and 7).

Given a Gaussian Mixture Model M obtained from training data

$$f(x) = \sum_{i=1}^{n} \alpha_i \mathcal{N}_{\mu_i,\sigma_i}(x) = \sum_{i=1}^{n} \alpha_i \mathcal{N}_M(x) \quad (13)$$

the energy $\mathcal{L}_M(x)$ for corresponding model M at point x is defined as

$$\mathcal{L}_M(x) = \min \left\{ \left(\frac{\mu_i - x}{\sigma_i} \right)^2 : i = 1 \dots n \right\} - \theta \quad (14)$$

where θ denotes a threshold parameter defining what square distance from mean value is considered acceptable. In our setup the value θ was set so that 95% values from training data is accepted.

Let E' denote a subset of edges $\{e \in E \,|\, f(e) = 1\}$ of the graph G and let $C(G, f)$ denote a set of strongly connected components of the graph G when taking into account only edges in E'.

The energy of character height ratios $\mathcal{E}_{hr}(G, f)$ is defined as

$$\mathcal{E}_{hr}(G, f) = \sum_{e \in E'} \mathcal{L}_{M_{hr}} \left(\mathsf{hr}\left(e_b, e_e\right) \right) \quad (15)$$

where e_b (e_e) denotes a vertex where the edge e begins (ends).

The energy of character centroid angles $\mathcal{E}_{ca}(G, f)$ is defined as

$$\mathcal{E}_{ca}(G, f) = \sum_{\substack{e^1, e^2 \in E' \\ e^1_e = e^2_b}} \mathcal{L}_{M_{ca}}(\mathsf{ca}(e^1_b, e^1_e, e^2_e)) \quad (16)$$

where again e^i_b (e^i_e) denotes a vertex where the edge e^i begins (ends).

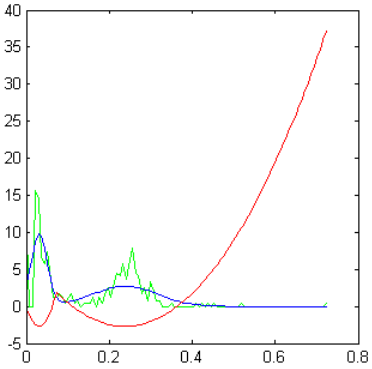


Figure 11. Normalized histogram of training data (green), inferred Gaussian Mixture Model M (blue) and corresponding energy function $\mathcal{L}_M$ (red)

The energy of text direction distances $\mathcal{E}_d(G, f)$ and energy of line approximation $\mathcal{E}_{la}$ are defined as

$$\mathcal{E}_d(G, f) = \sum_{cc \in C(G,f)} \sum_{c \in cc} \mathcal{L}_{M_d} \left(\frac{d(c, \tau)}{h_{\max}} \right) \quad (17)$$

$$\mathcal{E}_{la}(G, f) = \sum_{cc \in C(G,f)} \max \left\{ \mathcal{L}_{M_{tb}} \left(\frac{h_t(\tau)}{h_{\max}} \right), \mathcal{L}_{M_{mb}} \left(\frac{h_m(\tau)}{h_{\max}} \right), \mathcal{L}_{M_{bb}} \left(\frac{h_b(\tau)}{h_{\max}} \right) \right\} \quad (18)$$

$$\tau = \mathsf{la}(cc),\ h_{\max} = \max_{c' \in cc}(c'_b - c'_t)$$

3.3. Building region sequences

Region sequences are iteratively built by altering the graph configuration f in order to minimize the energy of the graph $\mathcal{E}(G, f)$. In each step the procedure `test` compares energy of newly created graph configuration f' to the best energy found so far and if a lower energy is found, the current configuration f is updated.

The method starts by enumerating all region triplets, taking only the acceptable triplets (the ones which decrease the graph energy $E(G, f)$) and thus initializing values of text line hidden parameters. Then, single regions are enumerated and the hidden text line parameters are used to efficiently prune the search space. As a last step the method tries to disconnect regions based on the inferred parameters of the whole line of text, because some regions might have been connected in the early stage as a result of inaccurate hidden parameters estimation on short sequences. The process is outlined in Figure 12, a result of the process is shown in Figure 13.

```
Procedure findBestConfiguration (G)
 f := (0,0, ... 0)
 E := 0
 { Connecting triplets of regions to obtain
   initial values of hidden parameters }
 for each subsequent pair of edges e,e' in G
  f' := f
  f'(e, e') = 1
  (E, f) := test(E, f, f')

 { Connecting single regions }
 for each edge e in G
  f' := f
  f'(e) := 1
  (E, f) := test(E, f, f')

 { Trying to disconnect pairs of nodes }
 for each edge e in G
  f' := f
  f'(e) := 0
  (E, f) := test(E, f, f')

 return f

Procedure test(E, f, f')
 E' = calculateEnergy(f')
 if E' < E
  E := E'
  f := f'

 return (E, f)
```

Figure 12. Pseudo-code of finding the best region graph configuration f in the region sequences building process

Figure 13. Region graph and its edge labeling corresponding to the best configuration (edges of the graph $f(e) = 1$ marked green, $f(e) = 0$ marked red)

4. Experiments

The method was evaluated using the hypothesis-verification framework proposed by Neumann and Matas [10] and replacing the heuristics text formation stage by the proposed method. The standard and most cited ICDAR 2003 Robust Reading Competition dataset[1] [8] was used for performance evaluation. The Train set was used to obtain the method parameters and an independent Test set was used to evaluate the performance. In total the ICDAR 2003 Test set contains 5370 letters and 1106 words in 249 pictures.

method	precision	recall	f
Pen et. al [11]	0.67	0.71	0.69
Zhang et. al [13]	0.73	0.62	0.67
Epshtein et. al [4]	0.73	0.60	0.66
Hinnerk Becker [7]	0.62	0.67	0.62
proposed method	**0.60**	**0.60**	**0.60**
Alex Chen [7]	0.60	0.60	0.58
Neumann and Matas [10]	0.59	0.55	0.57
Ashida [8]	0.55	0.46	0.50
HWDavid [8]	0.44	0.46	0.45
Wolf [8]	0.30	0.44	0.35
Qiang Zhu [7]	0.33	0.40	0.33
Jisoo Kim [7]	0.22	0.28	0.22
Nobuo Ezaki [7]	0.18	0.36	0.22
Todoran [8]	0.19	0.18	0.18

Table 1. Text localization results on the ICDAR 2003 dataset

Applying the evaluation protocol defined in [8], the proposed method achieved precision of 0.60 and recall of 0.60, which gives f-measure of 0.60. Figure 14 shows examples of text localization and recognition on the ICDAR 2003 dataset.

5. Conclusions

A novel method for text line formation was proposed. The method uses the hidden parameters of the text line (such as text direction) to group Maximally Stable Extremal Regions (MSERs) into lines of text. The exhaustive enumeration of short sequences is achieved by finding all character region triplets that fulfill constraints of textual content, which keeps the proposed method efficient yet still capable to perform a robust estimation of the hidden parameters in order to correctly initialize the search.

The proposed method was evaluated on the standard ICDAR 2003 dataset using the standard evaluation protocol [8], where it outperforms the method for forming text lines of Neumann and Matas [10] (f-measure is increased from 0.57 to 0.60). The method is still behind the state-of-the-art method for text localization (Pen et al. [11], f-measure 0.69), but the text localization results have to be interpreted carefully as there are known problems with the evaluation

[1] http://algoval.essex.ac.uk/icdar/Datasets.html

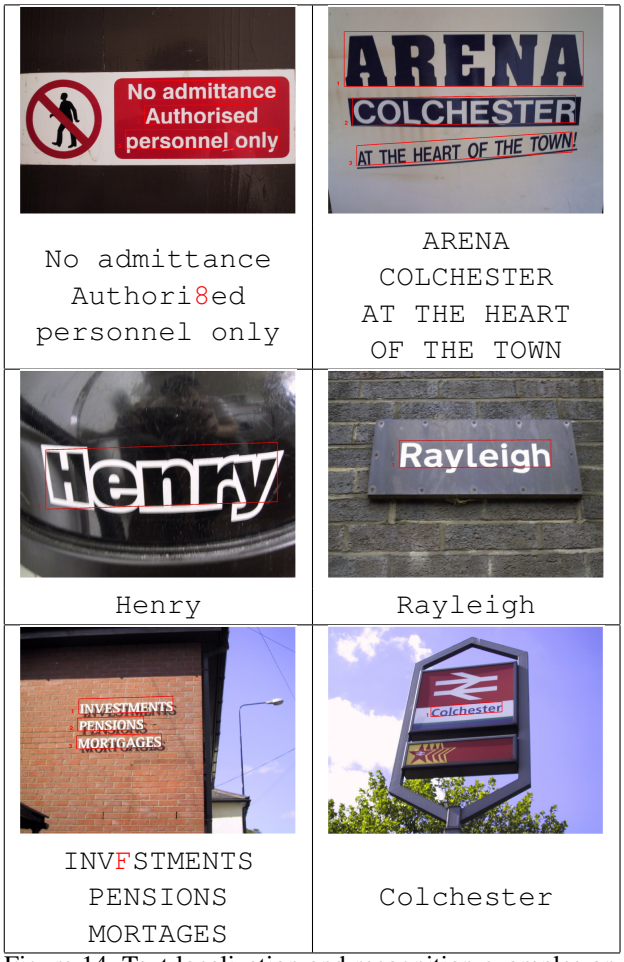


Figure 14. Text localization and recognition examples on the ICDAR 2003 dataset.

protocol and ground truth of the ICDAR 2003 dataset [7, 10]. The proposed method aims to solve the complete problem of text detection and recognition (see Figure 14), however all the methods superior in text localization performance [11, 13, 4, 7] aim only to solve one part of the problem and thus direct comparison cannot be made.

Most frequent problems of the proposed method is unsupported text line structure (Figure 15a), symbols or pictographs placed close to text lines (Figure 15b), letters not detected as individual regions (Figure 15c) and false positives caused by repetitive textures with a text-like spacial structure (Figure 15d).

Acknowledgement. The authors were supported by Czech Government reseach program MSM6840770038.

(a) (b) (c) (d)

Figure 15. Problems of the proposed method. (a) Unsupported text line structure. (b) Pictographs placed close to text lines. (c) Letters not detected as individual regions. (d) False positives caused by repetitive textures with a text-like spacial structure

References

[1] X. Chen, J. Yang, J. Zhang, and A. Waibel. Automatic Detection and Recognition of Signs From Natural Scenes. *IEEE Trans. on Image Processing*, 13:87–99, Jan. 2004. 1

[2] X. Chen and A. L. Yuille. Detecting and reading text in natural scenes. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:366–373, 2004. 1

[3] M. Donoser, H. Bischof, and S. Wagner. Using web search engines to improve text recognition. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1 –4, 2008. 1

[4] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *CVPR '10: Proc. of the 2010 Conference on Computer Vision and Pattern Recognition*, 2010. 1, 6, 7

[5] N. Ezaki. Text detection from natural scene images: towards a system for visually impaired persons. In *In Int. Conf. on Pattern Recognition*, pages 683–686, 2004. 1

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

[7] S. M. Lucas. Text locating competition results. *Document Analysis and Recognition, International Conference on*, 0:80–85, 2005. 6, 7

[8] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Icdar 2003 robust reading competitions. In *ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition*, page 682, Washington, DC, USA, 2003. IEEE Computer Society. 1, 6

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable

extremal regions. *Image and Vision Computing*, 22(10):761–767, September 2004. 1, 4

[10] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In R. Kimmel, R. Klette, and A. Sugimoto, editors, *ACCV 2010: Proceedings of the 10th Asian Conference on Computer Vision*, volume IV of *LNCS 6495*, pages 2067–2078, Heidelberg, Germany, November 2010. Springer. 1, 4, 6, 7

[11] Y.-F. Pan, X. Hou, and C.-L. Liu. Text localization in natural scene images based on conditional random field. In *ICDAR '09: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*, pages 6–10, Washington, DC, USA, 2009. IEEE Computer Society. 1, 6, 7

[12] V. Wu, R. Manmatha, and E. M. Riseman, Sr. Textfinder: An automatic system to detect and recognize text in images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(11):1224–1229, 1999. 1

[13] J. Zhang and R. Kasturi. Character energy and link energy-based text extraction in scene images. In R. Kimmel, R. Klette, and A. Sugimoto, editors, *ACCV 2010: Proceedings of the 10th Asian Conference on Computer Vision*, volume II of *LNCS 6495*, pages 832–844, Heidelberg, Germany, November 2010. Springer. 6, 7