Unconstrained Licence Plate and Text Localization and Recognition

Jiri Matas (*matas@cmp.felk.cvut.cz*)
Karel Zimmermann (*zimmerk@cmp.felk.cvut.cz*)
Czech Technical University Prague
Center for Machine Perception *http://cmp.felk.cvut.cz*

Abstract

Licence plates and traffic signs detection and recognition have a number of different applications relevant for transportation systems, such as traffic monitoring, detection of stolen vehicles, driver navigation support or any statistical research. A number of methods have been proposed, but only for particular cases and working under constraints (e.g. known text direction or high resolution).

Therefore a new class of locally threshold separable detectors based on extremal regions, which can be adapted by machine learning techniques to arbitrary shapes, is proposed. In the test set of licence plate images taken from different viewpoints $\langle -45^o, 45^o \rangle$, scales (from seven to hundreds of pixels height) even in bad illumination conditions and partial occlusions, the high detection accuracy is achieved (95%). Finally we present the detector generic abilities by traffic signs detection.

The standard classifier (neural network) within the detector selects a relevant subset of extremal regions, i.e. regions that are connected components of a thresholded image. Properties of extremal regions render the detector very robust to illumination change and partial occlusions. Robustness to a viewpoint change is achieved by using invariant descriptors and/or by modelling shape variations by the classifier.

The time-complexity of the detection is approximately linear in the number of pixel and a non-optimized implementation runs at about 1 frame per second for a 640×480 image on a high-end PC.

I. Introduction

Text detection and recognition in general, and licence plate detection and recognition in particular, in images and videsequences have a number of different applications relevant for transportation systems, such as [6], [5] traffic monitoring, detection of stolen vehicles, driver navigation support etc. Localisation (detection) is typically the critical problem [6], since commercial systems for recognition of printed text are available (e.g. [1]). Published methods such as operate under rather restrictive constraints, e.g. the orientation of the text (licence plate) is known. Ezaki [5] describes three methods of automatic detection based on morphogical operations and edge detection, but region of interest have to be usually manually choosen and the text is expected to be viewed frontally. In [2], Chen et al. propose an automatic text detection method for moving motor vehicles, but it assumes that region of interest and the text direction is given by movement of vehicle. The partially affine and illumination invariant detection method based on morphological operations is proposed in [3] but the experiments include only images with closed-look with high resolution of LP (character height about 50 pixels) and the view invariance is presented in a quite limited range.

As a main contribution of the paper, we propose a general approach for text and licence plate detection. By general we mean: insensitive to geometric changes induce by the change of viewpoint (scaling, rotation, affine deformation), illumination insensitive and robust to occlusion. The detector is based on the concept of *extremal regions* [4], which are arbitrary threshold-separable region.

In presented experiments (Section III) 95% detection accuracy is achieved on licence plate dataset in

- the horizontal $\langle -45^o, 45^o \rangle$ and the vertical $\langle -30^o, 30^o \rangle$ range of viewpoints, respectivelly and
- scales from the height of 7 pixels up to height appropriate to the image rows and
- illumination changes from over-illuminated images to the night images and
- even partially occluded LPs by hand or by other cars.

In our work, the objects of interest (e.g. sign) is decomposed into the spatial configuration of category-specific extremal regions (e.g. letters), i.e. a subset of extremal regions that is selected by machine learning methods and is likely to correspond to a letter or digit. In the detection stage, we first independently detect all character-like extremal regions and the longest linear spatial configuration of these region is labeled as text or licence plate. Further processing such as character recognition follows.

A robust category-specific detector of extremal regions can be implemented as follows. Enumerate all extremal regions, compute efficiently a description of each region and classify the region as relevant or irrelevant. In a learning stage, the classifier is trained on examples of regions – letters from given category (i.e. font, digits, capital etc.). Such detection algorithm is efficient only if features (descriptors) for each region are computed in constant time. We show there is a sufficiently discriminative class of 'incrementally computable' features on extremal regions satisfying this requirement.

The features are only scale invariant, the viewpoint invariance in defined range is achieved by training the classifier for different views from the given range. This fact let the detector

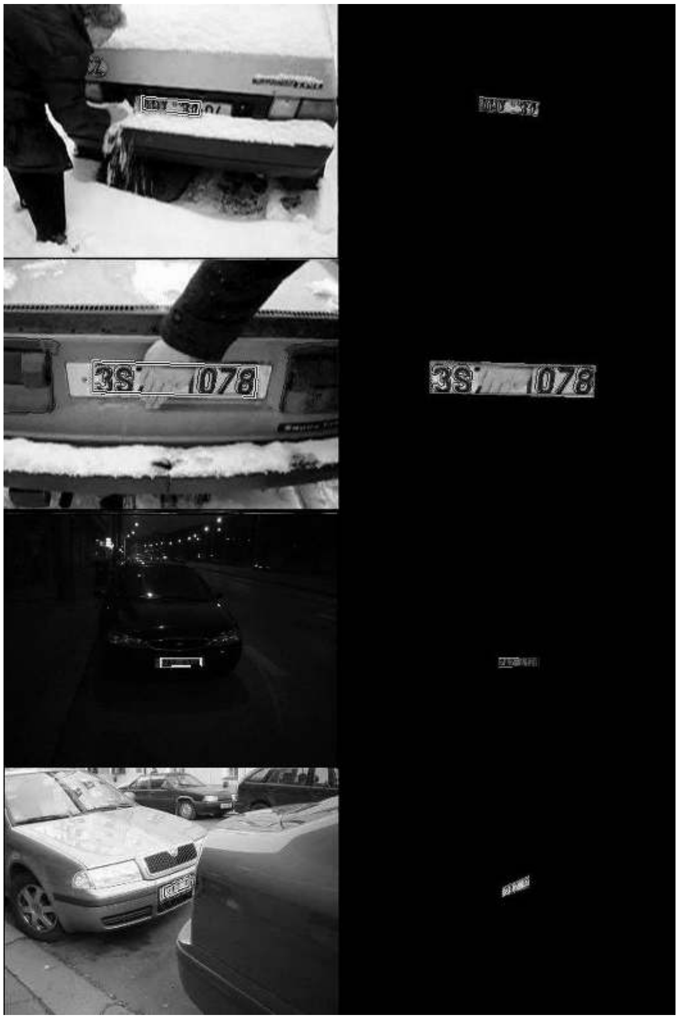

Fig. 1. Licence plate detection in unconstrained conditions.

to be trained for special cases with high accuracy of detection or to be trained for general text detection. In a special case like detection of licence plates on the car waiting in front of the bar, the classifier can be trained only for such special position of licence plate to the camera and appropriate font of capital letters and digits. The detection in more general cases is limited by ability of classifier to conceive the number of different examples. Therefore accuracy is decreasing with generalization of the problem.

The rest of the paper is organised as follows. First, the structure of the algorithm for category-specific extremal region detection is presented. We show that CSER are efficiently selected by interleaving enumeration of extremal regions and classification of their incrementally computable features. The class of incrementally computable features is studied next, necessary conditions for the class are found and examples of such features are given (Section II-A). We than apply the method to two well known problems of text detection and license plate recognition (Section III). The paper is summarised in Section IV.

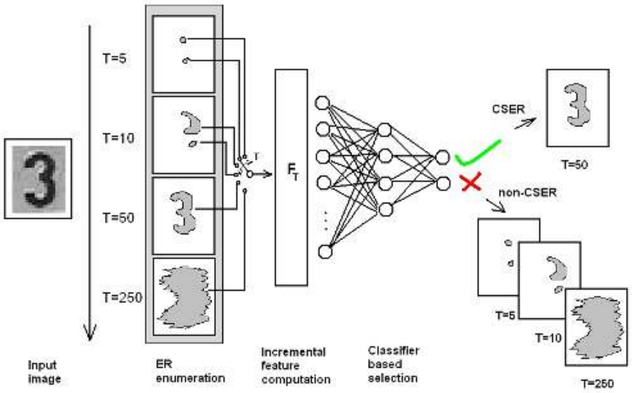

Fig. 2. The detection is implemented as interleaved enumeration of extremal regions, computation of incremental features and classification.

II. Category-specific extremal region detection

Our objective is to select from the set of extremal regions those with shape belonging to a given text category. The model of the text is acquired in a separate training stage. Let us assume for the moment that the learning stage produced a classifier that, with some error, is able to assign to each extremal region one of two labels: 'interesting', i.e. is a component of our category, or 'non-interesting' otherwise. The detection of category-specific extremal regions can be then arranged as three interleaved steps: (1) generate a new extremal region, (2) describe the region and (3) classify it. The interleaved computation is schematically depicted in Figure 2.

Extremal regions are connected components of an image binarised at a certain threshold. More formally, an extremal region r is a contiguous set of pixels such that for all pixels $p \in r$ and all pixels q from the outer boundary ∂r of region r either $I(p) < I(q)$ or $I(p) > I(q)$ holds. In [4], it is shown that extremal regions can be enumerated simply by sorting all pixels by intensity either in increasing or decreasing order and marking the pixels in the image in the order. Connected components of the marked pixels are the extremal regions. The connected component structure is effectively maintained by the union-find algorithm.

In this process, exactly one new extremal region is formed by marking one pixel in the image. It is either a region consisting of a single pixel (a local extremum, a region formed by a merge of regions connected by the marked pixel, or a region that consisting of union of an existing region and the marked pixels. It is clear from this view of the algorithm that there are at most as many extremeral regions as there are pixels in the image. The process of enumeration of extremal regions is nearly linear in the number of pixels [1] and runs at approximately 10 frames per second on 2.5 GHz PC for a 700×500 image.

To avoid making the complexity of the detection process quadratic in the number of image pixels, the computation

[1] The (negligibly) non-linear term is hidden in the "maintenance of connected component structure".

of region description must not involve all of its pixels. Fortunately, a large class of descriptors can be computed incrementally in constant time even in the case of a merge of two or more extremal regions (the other two situations are special cases). Importantly, combinations of incrementally computable features include affine and scale invariants. Incrementally computable features are analysed in Section II-A.

The final step of the CSER detection, the selector of category-specific regions, is implemented as a simple classifier (NN, adaboost) trained on examples of regions - components of the category of interest. The classifier selects relevant regions in constant time. The overall process of marking a pixel, recalculating descriptors and classifying is thus constant time. The choice of classifier is arbitrary and any other classifier such as SVM could replace it.

At this point it is interesting to compare the proposed CSER detection process with the seminal face detection method of Viola and Jones [7]. Viola and Jones use cascaded Adaboost to classify (in principle) every rectangular window of a predetermined size using features computed in constant time from the integral image. There are strong analogies. In both cases, the number of classifications made is equal to the number of pixels in the image (which is equal both to the number of rectangular windows of fixed and the number of extremal regions). In both cases, features describing the classified regions are computed in constant time. In the Viola-Jones approach the assumption is that the object from the category (faces) are well represented in a rectangular window. In our case, the assumption is that the category of interest has components that are extremal regions. The difference in the adopted classifier is superficial.

A. Incrementally Computable Region Descriptors

In the CSER detection process we are given two or more disjoint regions r_1 and r_2. By marking a pixel in the image, these regions merge to form a new extremal region. The new region is the union of $r_1 \cup r_2$ (we use $r_.$ to identify both the region and its set of pixels). The following problem arises: what image features computed on the union of the regions can be obtained in constant time from some characterisation g of r_1 and r_2?

For example, let us suppose that we want to know the second central moment of the merged region. It is known that the second central moment (moment of inertia) can be computed in constant time from the first and second (non-central) moments and first and second (non-central) moments can be updated in the merge operation in constant time. A region descriptor (feature) ϕ will be called *incrementally computable* if the following three functions exists: a characterising function $g : 2^{Z^2} \rightarrow \mathcal{R}^m$, a characterisation update function $f : (\mathcal{R}^m, \mathcal{R}^m) \rightarrow \mathcal{R}^m$, and a feature computation function $\phi : \mathcal{R}^m \rightarrow \mathcal{R}^n$, where m is constant, n is the dimension of the feature and Z^2 is the image domain.

For each region, the characterising function g returns the information necessary for computing feature ϕ in a real vector of dimension m. The dimension m of the characteristic vector depends on the feature, but is independent of region size. Given the characterisation returned by g, the n-dimensional feature of interest (region descriptor) is returned by ϕ. Function f computes the characterisation of the merged region given the characterisation of the regions r_1, r_2. For efficiency reasons, we are looking for features with the smallest characterisation dimension m^*. An incremental feature is a triplet of functions (g^*, f^*, ϕ^*) defined as

$$g^* = \arg\min_g \{\dim(g(2^{Z^2}))\} \text{ subject to}$$

$$\phi(g(r_1 \cup r_2)) = \phi(f(g(r_1), g(r_2))).$$

Example 1. Minimum intensity I of all pixels in a region is an incrementally computable feature with dimension $m^* = 1$. Given regions r_1 and r_2 with pixels $r_1^i \in r_1, r_2^j \in r_2$, the description of the union regions r_1, r_2 is

$$\phi(g(r_1 \cup r_2)) = \underbrace{1}_{\phi} \cdot \underbrace{\min}_{f} \{ \underbrace{\min_{r_1^i \in r_1} I(r_1^i)}_{g(r_1)}, \underbrace{\min_{r_2^j \in r_2} I(r_2^j)}_{g(r_2)} \}$$

Example 2. The center of gravity ($m^* = 2$) of a union of regions r_1, r_2 with pixels r_1^i, r_2^j for $i = 1...k_1, j = 1...k_2$ is

$$\phi(g(r_1 \cup r_2)) = \underbrace{\frac{1}{k_1 + k_2}}_{\phi} \left(\underbrace{\sum_{i=1}^{k_1} r_1^i}_{g(r_1)} \underbrace{+}_{f} \underbrace{\sum_{j=1}^{k_2} r_2^j}_{g(r_2)} \right).$$

In this paper we use the following incrementally computable features: *normalized central algebraic moments* with $m^* \sim (k)^2$ where k is an moment order (calculation based on algebraic moments), *compactness* with $m^* = 2$ (using the area and the border), *Euler number* of a region with $m^* = 2$, *Entropy of cumulative histogram* with $m^* = 2$. Features that we are not able to compute incrementally are e.g. the number convexities and the area of convex hull.

All of the extremal regions can be enumerated and classified in linear time with respect to number of point in image. The algorithm **A1** of fast feature enumeration is shown below. $\mathcal{I}$ is the image and $\mathcal{R}, \mathcal{R}^*$ are set of current extremal regions (appropriate to the threshold T) and set of CSER (subset of extremal regions), respectively.

```
A1 - The algorithm of the CSER detection
  1) $T = 1$, $\mathcal{R}^* = \emptyset$ and $\mathcal{R} = \emptyset$.
  2) For all pixels $p$ such that $\mathcal{I}(p) = T$
       a) Append pixel - If $\exists! r_i \in \mathcal{R},\ i = 1..n,\ p \in \partial r_i$ then $r_i = r_i \cup p$.
       b) Merge regions - If $\exists r_i, r_j \in \mathcal{R}, i, j = 1..n, i \neq j, p \in \partial r_i \wedge p \in \partial r_j$ then $r_i = r_i \cup r_j \cup p \wedge \mathcal{R} = \mathcal{R} \backslash r_j$.
       c) New region - If $\neg \exists r_i \in \mathcal{R}, i, j = 1..n, p \in \partial r_i$ then $r_{n+1} = p,\ \mathcal{R} = \mathcal{R} \cup r_{n+1}$.
  3) For each new or changed regions $r \in \mathcal{R}$ recompute features. The regions $r_i$ with positive classification are CSER
         $\mathcal{R}^* = \mathcal{R}^* \cup r_i$
  4) If $T < T_{max}$ then $T = T + 1$ and continue at 2, else end.
```

(a)

$\theta \backslash \phi$	0^o	15^o	30^o	45^o
0^o	2.6	2.8	2.8	3.0
10^o	3.2	3.2	3.2	3.8
20^o	3.2	3.6	4.0	7.8
30^o	7.6	8.4	15.2	26.5

(b)

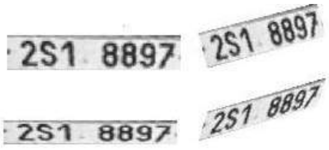

Fig. 4. (a) False negative rate (missed CSER on licence plates) as a function of viewing angles ϕ (elevation) , θ (azimuth); in percentage points. (b) An Example of a syntheticaly warped licence plate to ϕ, θ equal to $(0^o, 0^o), (0^o, 45^o), (30^o, 0^o)$ and $(30^o, 45^o)$.

III. Experiments - Applications of CSER detection

A. Licence plate detection

At least in constrained condition, licence plate detection, as demonstrated e.g. by the London congestion charge system, is more an engineering than a research problem. Here we demonstrate that an unconstrained licence plate detector is developed easily (and without ad hoc tricks) using CSERs. By 'unconstrained detector' we mean viewpoint and illumination independent and robust to occlusion.

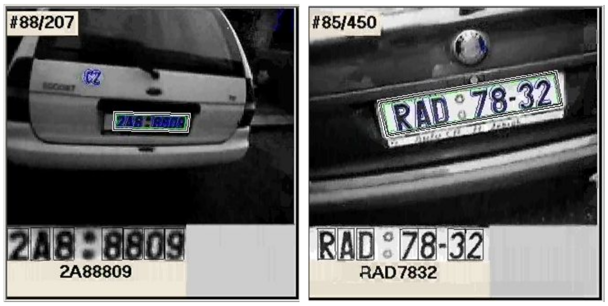

Fig. 3. Licence plate detection and recognition. Detected CSER are highlighted by blue color, the appropriate longest linear configuration is highlighted by green color. Then the detected LP is normalized and recignized by standart OCR method.

The category of licence plates is modelled as a linear constellation of CSERs. Information about the rectangular shape of the place as a whole is not exploited. The feed-forward neural network for CSER selection was trained by a standard back-propagation algorithm on approximately 1600 characters semi-automatically segmented from about 250 images acquired in unconstrained conditions. The region descriptor was formed by scale-normalised algebraic moments of the characteristic function up the fourth order, compactness and entropy of the intensity values. Intentionally, we did not restrict the features to be either rotation or affine invariant and let the neural network with 15 hidden nodes to model feature variability. Counterexamples were obtained by ten rounds of bootstrapping. In each round, he CSER detector processed the 250 training images and the false positives served as negative examples in the next round of training.

The detection of licence plates proceeds by in two steps. First, relevant CSER selected as described above. Second, linear configurations of regions are found by Hough transform. We impose two constraints on the configurations: the CSER regions must be formed from more than three regions and the regions involved must have a similar height.

Detection Rate. On an independent test set of 70 unconstrained images of scenes with licence plates the method achieved 98% detection rate with a false positive appearing in approximately 1 in 20 images. Example of detected licence plates and the type of data processed are shown in Figure 1.

Speed. The detection time is proportional to the number of pixels. For a 2.5 GHz PC the processing took 1.1 seconds for a 640×480 image and 0.25 seconds for 320×240 image.

Robustness to viewpoint change was indirectly tested by the large variations in the test data where scales of licence plates differed by a factor of 25 (character 'heights' ranged from approximately 7-8 to 150 pixels) and the plates were viewed both frontally and at acute angles. We also performed systematic evaluation of the CSER detector. Images of licence plates were warped (see Figure 4b) to simulate a view from a certain point on the viewsphere. The false negative rates of the CSER detector (with approximately 10 false positive regions per street background image) are shown in Table 4(a). The CSER detector is stable for almost the whole tested range. Even the 27% false negative at the 30^o-45^o elevation-azimuth means that three quarters of characters on the licence plate are detected which gives high probability of licence plate detection.

Robustness to illumination change was evaluated in a synthetic experiment. Intensity of images taken in daylight was multiplied by a factor ranging from 0.02 to 1. As shown in Figure 6, the false negative (left) and false positive (right) rates were unchanged for both the detector of CSER (bottom) and whole licence plates (top) in the $(0.1, 1)$ range! For the 0.1 intensity attenuation, the image has at most 25 intensity levels, but thresholds still exist that separate the CSERs. Moreover, if the region is well threshold separable (i.e. we explicitly know that there is a wide scale of thresholds which separate object from background) then we can add a constraint of detection stability. This fact was not used in any of these experiments to demonstrate the power of method. Therefore we are able to detect LPs in the night scenes even partially overlighted by light sources in Fig.5b The classifier output for character-like region and non-character region is shown in Fig.5a.

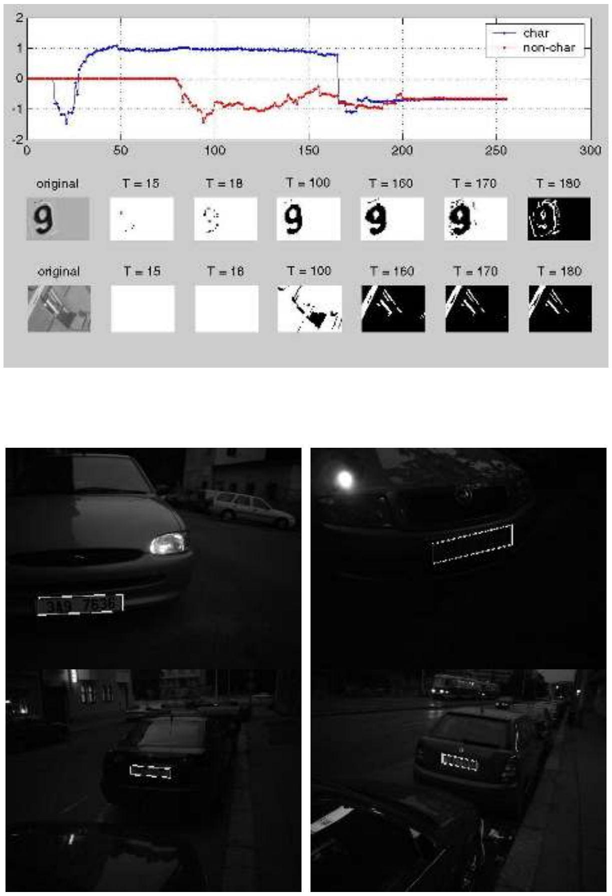

Fig. 5. (a) The classifier output for character-like region and non-character region. The Figure shows that classification of letter is robust to illumination change. The growing region is classified as character over a very large range of thresholds (40–160). (b) Licence plate detection in night shows brightness invariance of proposed method.

Robustness to occlusion as demonstrated in Figure 1b is a consequence of modelling the object as a configuration of local component. Occlusion of some components does not imply the object is not detected.

Another (almost same) application of CSER detection arises from segmentation of letters in detected and normalized licence plates 7. The LPs are detected even if a few letters missing, but we are looking for the complete segmentation which is input of optical character recognition (OCR). In a similar way, we use CSER detector trained on normalized letters (i.e. the build-in classifier is trained only for letters in frontal position). In the segmention problem we achieved FP=FN=1.6% accuracy.

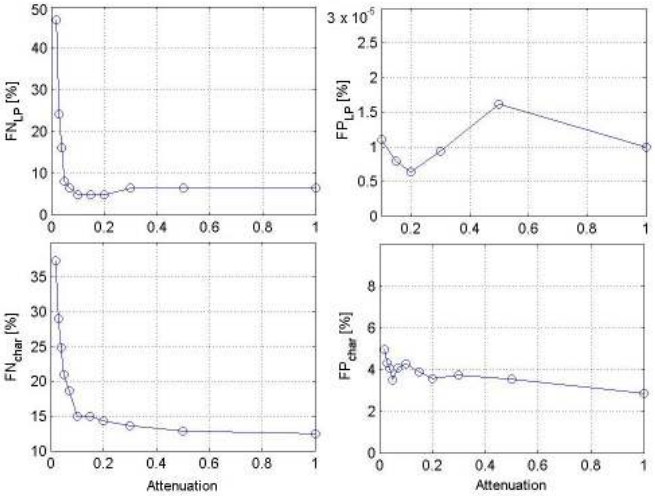

Fig. 6. Licence plate detection in images with attenuated intensity.

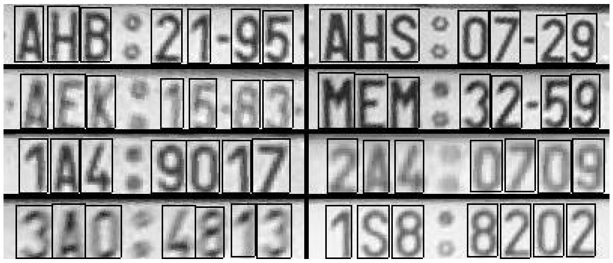

Fig. 7. Licence plate characters segmentation via CSER

B. *Traffic signs detection by thresholding in the given direction*

In this experiment we outlined the way how to use the proposed detector for color images. The extremal regions are defined for scalar function of a total ordering of pixels. In the case of gray level image the scalar function is the intensity of pixels in the case of color image the scalar function is $\lambda : \mathcal{R}^3 \to \mathcal{R}$ which assigns the scalar value for each triplet of RGB components.

In the task of traffic signs detection we defined total ordering by

$$\lambda(RGB) = \frac{1 - R + B}{2},$$

where R and B are components of unit color vector (red and blue).

The results and the scalar function transformation of original image are shown in Fig.8. In this task we do not have enough regions for post-filtering of FP in a way of linear constellation constraints. Therefore we present the detector as rapid pre-selector with small value of FN.

Again in this experiment we notice both brightness and scale invariance of the detector. The rotation invariance is provided by training of internal classifier on the rotated examples. Moreover we note presence of blur robustness in comparism with other approaches due to the fact that extremality of the region is not usually significantly changed by bluring.

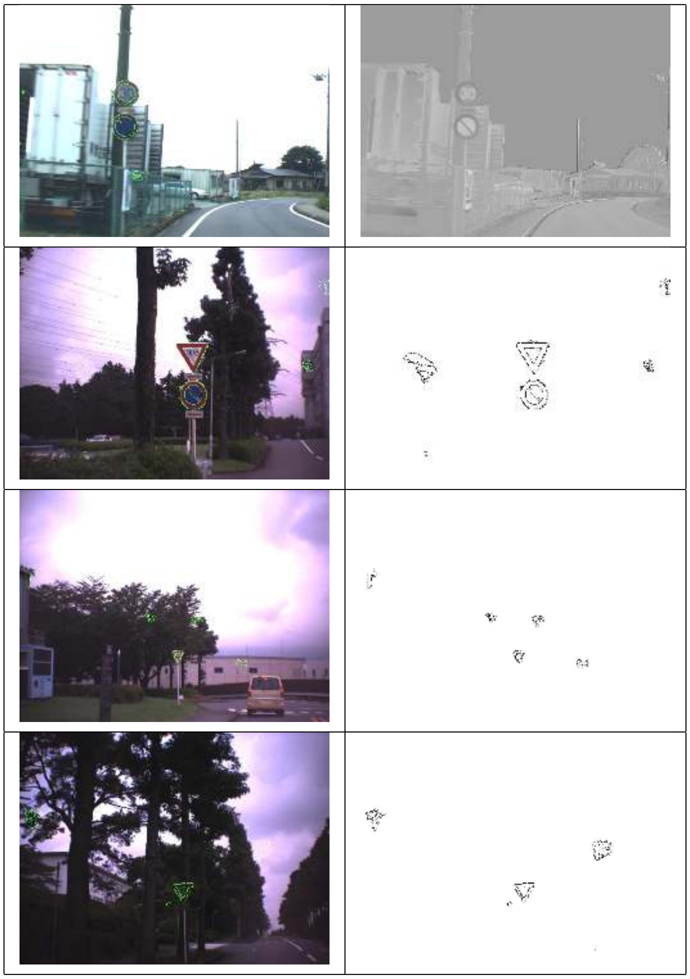

Fig. 8. Traffic signs detection results: First row shows traffic signs original image, then image thresholded in RB direction (i.e. from red to blue color) and detected regions. Other rows shows different images and the detected regions.

In future experiments we will show that the direction of thresholding can be object of machine learning methods too (i.e. that we are able to learn the transformation from color to threholdable image).

IV. Conclusions

We presented a detector that can be adapted by machine learning methods to detect arbitrary locally threshold-separable regions from a given category (e.g. letters, signs). The detector selects a category-relevant subset of extremal regions. Properties of extremal regions render the detector very robust to illumination change. Robustness to viewpoint change can be achieved by using invariant descriptors and/or by modelling shape variations by the classifier.

We showed high effectivity of proposed method in the problem of robust, view point and scale invariant licence plate detection. The training and test views included 25-fold change of scale and views form acute angles. We used approximately 250 training images (i.e. 1600 characters). Results have been verified on testing set which consists of 70 licence plates. We missed only one licence plate and three plates where found in the image without any plates (e.g. bushes). In the other words, false negative rate was 1.5% and false positive rate was 4%. These results were achieved without any post-processing based on character recognition. Additionaly, we used the detector for character segmentation from normalized licence plates with FP=FN=1.5%.

The experiment of traffic signs detection presents the proposed detector application for color images. An interesting results were shown in the experiments but we noted high rate of false positive due to the number of different fonts. The false positive rate could be decreased by additional suceeding region filtering by strong classifier.

In the future work we want to show that direction of threholding could be find by methods of machine learning, i.e. that we are able to learn the colors of regions we are looking for and use it for such thresholding which locally separate them from their backgrounds and make them detectable in the presented way like CSER.

The time-complexity of the detection is approximately linear in the number of pixel and the current implementation runs at about 1 frame per second for a 640×480 image on a high-end PC.

Acknowledgements

Authors was supported by The Austrian Ministry of Education under project CONEX GZ 45.535 and by The Czech Academy of Sciences under project 1ET101210407.

References

[1] ABBYY. Abbyy finereader 7.0 corporate edition. In *http://buy.abbyy.com/content/frce/default.aspx*, 2005.

[2] Datong Chen, Jean-Marc Odobez, and Herve Bourlard. Text Detection and Recognition in Images and Videos. *Pattern Recognition*, 37(3):595–609, March 2004.

[3] Jun-Wei Hsieh, Shih-Hao Yu, and Yung-Sheng Chen. Morphology-based license plate detection from complex scenes. In *16 th International Conference on Pattern Recognition (ICPR'02)*, volume 3, pages 176–180, 2002.

[4] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC02*, volume 1, pages 384–393, London, UK, 2002.

[5] N. Nobuo Ezaki, M. Bulacu, and L. Schomaker. Text detection from natural scene images: Towards a system for visually impaired persons. In *International Conference on Pattern Recognition*, volume 2, pages 683–686, Cambridge, UK, 2004.

[6] M. Shridhar and J. Miller. Recognition of license plate images: Issues and perspectives. In *Fifth International Conference on Document Analysis and Recognition*, pages 17–21, 1999.

[7] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision - to appear*, 2002.