ISSN 1392–124X (print), ISSN 2335–884X (online) INFORMATION TECHNOLOGY AND CONTROL, 2015, T. 44, Nr. 1

Scene Text Extraction in IHLS Color Space Using Support Vector Machine

Matko Šarić

Faculty of Electrical Engineering, Mechanical Engineering and Naval Arhitecture
University of Split R. Boskovica 32, Split Croatia
e-mail: msaric@fesb.hr

Hrvoje Dujmić

Faculty of Electrical Engineering, Mechanical Engineering and Naval Arhitecture
University of Split R. Boskovica 32, Split Croatia
e-mail: hdujmic@fesb.hr

Mladen Russo

Faculty of Electrical Engineering, Mechanical Engineering and Naval Arhitecture
University of Split R. Boskovica 32, Split Croatia
e-mail: mrusso@fesb.hr

crossref http://dx.doi.org/10.5755/j01.itc.44.1.5757

Abstract. Scene text extraction is a challenging research area due to variety of image degradations caused by imaging conditions and low cost consumer devices. In this paper we propose a text extraction method that uses chroma and lightness component for generation of extraction hypotheses and incorporates SVM (support vector machine) based text detection stage as tool for hypotheses verification. The choice of chroma and lightness components is based on their complementarity with respect to image degradations like shadows and highlights. Another novelty is the usage of IHLS color space for text extraction task which is motivated by saturation definition that eliminates instability of this component at low lightness values. Results obtained on the ICDAR 2011 dataset confirm complementarity of chroma and lightness. Compared to the state-of-the-art methods, the proposed algorithm achieves higher correct recognition rate and comparable total edit distance.

Keywords: scene text extraction; IHLS color space; SVM; text detection.

1. Introduction

Text information extraction in images and video represents a challenging research area. In conjunction with increasing availability of low cost camera devices this field opens range of new applications as mobile language translator[1], text reader for visually impaired people [2], mobile visual search [3] etc. Text can be classified as text in documents, caption text and scene text [4]. Recognition of text in documents today achieves satisfactory results thanks to properties as regular layout, uniform background, good contrast etc. Caption text refers to artificially overlaid characters in images or video frames and it plays important role in content retrieval and indexing. Scene text is an integral part of image or video frame that can be found in our environment without constraints typical for document text. Growing popularity of cameras embedded in smartphones enables easy acquisition of scene text, but this kind of devices also introduces new imaging conditions as sensor noise, viewing angle, blur, lightning, resolution and aliasing [5]. Further, scene text is often characterized by non-uniform background, unknown layout, variety of fonts; it can be printed on materials with different reflection characteristics or on non-planar objects causing additional deformation. Combination of these circumstances increases complexity of scene text extraction and recognition. The problem of text information extraction consists of 5 subproblems [4]: detection, localization, tracking, extraction and

enhancement, and recognition. Some authors also divide this process in 3 steps: text localization, text segmentation and recognition of segmented characters[6,7]. Because of system complexity the authors usually address one of the previously mentioned steps in processing chain. This paper deals with step of scene text extraction, often referred as text segmentation or text binarization. The aim of the text extraction is separating character pixels from those of background. Input to this step is result from text localization stage, that is cropped image region containing text as major part. Text extraction directly influences on the success of recognition, especially in the case of scene text where imaging conditions and complex backgrounds make this task more challenging than in the case of document text. In this paper we propose a new scene text extraction method that firstly generates two hypotheses, that is two segmentations done with Otsu thresholding of chromatic distance from background color and lightness component in IHLS color space. The main assumption is that these components are complementary with respect to different degradations of scene text and it is confirmed experimentally. Chromatic distance representing hue and saturation is robust to shadows and highlights and it enables correct character segmentation in cases when the lightness component fails. This idea is inspired by text extraction method presented in [8] that exploits complementarity of angle distance, representing chromatic difference in RGB color space, and Euclidean distance representing intensity difference in RGB color space. We also propose the usage of IHLS space [9] in text extraction to avoid problems arising from instability of saturation component at low intensity levels. Next step is hypotheses verification: for each segmentation result, SVM (support vector machine) classifier is used to detect characters, that is to estimate "character − similarity" that tells us to what degree each connected component (CC) represents character or non-character. Extraction result with higher average "character − similarity" is chosen as correct extraction and further sent to recognition stage performed by OCR software. Our approach integrates connected component based text detection with text extraction step in order to choose better segmentation. In [7], multiple extractions are generated by dichotomization of subimages from K-means results, while in our work we reduce the number of assumptions using their complementarity. In [8], proper extraction is chosen using recognition results; contrary, in our method this step is performed by SVM-based text detection avoiding need for values of recognition performance. Obtained results justify usage of chroma and lightness for generation of hypotheses in our scene text extraction algorithm. Compared to the state-of-the-art methods, our algorithm achieves higher correct recognition rate and comparable total edit distance. The rest of the paper is organized as follows. Section 2 presents an overview of the related work. Section 3 describes the proposed method. Results are discussed in Section 4 and paper is concluded in Section 5.

2. Related work

In [5], text extraction methods are classified as thresholding-based and grouping-based. The first category includes global, local and entropy-based thresholding, while region-based, learning-based and clustering based-methods belong to the second group. Global thresholding uses histogram bimodality for separating characters from the background. In [2], text is extracted using Otsu threshold and global threshold calculated from mean intensities of character skeletons. Despite low computational requirements, this method often fails in case of scene text because of absence of obvious peaks in histogram. In adaptive or local binarization, threshold is defined locally for given image part based on statistical features of pixel values. Niblack method [10], where threshold is calculated based on local mean and deviation, is used in [11] showing high efficiency and robustness to image degradation. Zhu et al. [12] proposed non-linear Niblack binarization where they added two ordered statistics filters. In grouping-based methods, groups of text pixels are formed according to certain criteria. Region growing and split-and-merge algorithms, representing bottom-up and topdown strategies, belong to region based approaches. Although these methods are computationally expensive and dependent on parameter values, their advantage is integration of spatial information which plays important role in character extraction. Karatzas and Antonacopoulos [13] segment text in web images using split-and-merge technique for chromatic and achromatic image regions. These regions are recursively split using intensity histogram for achromatic pixels and hue histogram for chromatic pixels. In leaf layers of the resulting tree structure, the connected components are identified and further combined through merging process in order to extract characters. In [14], a 4-neighborhood region growing algorithm with Euclidean distance in RGB color space is used for background separation. In [15], Dujmic et al. proposed scene text extraction method employing region growing algorithm in HSI color space with modified cylindrical distance as homogeneity criterion. Seed pixels are selected based on horizontal projection of color differences between pixels and background. Learning-based methods employ classifiers like multi-layer perceptrons and support vector machines. This approach is more often used in text localization where classifier estimates text probability based on region feature vector [14]. Taking into account variety of scene text it is hard to build representative training database and it is the main reason why this kind of methods aren't employed more frequently in text extraction. Clustering-based methods rely on hypothesis that text and background pixels tend to form groups in an appropriate color space. The most popular technique is k-means although other clustering

approaches like GMM and spectral clustering also attract researcher's attention. Garcia and Apostolidis [16] segment text using 4-means algorithm in HSV color space. After background color estimation, vertical profile periodicity is employed to classify generated binary images as text or non-text. Mancas-Thillou and Gosselin [8] proposed text extraction method that uses clustering in RGB color space with two metrics: Euclidean distance and cosine similarity. Better output is selected using recognition results and it is further combined with Log- Gabor filtering of intensity component in order to exploit spatial information. In [7], k-means clustering is performed in HSI color space to generate a set of binarized images using dichotomization of K clusters. After segmentation of each binarized image in single character images, support vector machine is employed for estimation of character-likeness. Binarization with highest character string likeness is chosen as final result. Interconnection of different steps in text information extraction processing chain is presented in [6]: subtasks (text localization, character segmentation, text line formation) work with multiple hypotheses that are verified with feedback loops. Character/non-character classification is performed with SVM classifier and scale invariant connected component features. In [17], a method for container code detection is based on connected component features that produce good discrimination between characters and other objects. Classification is done using cascade of regularized least squares classifiers. For character detection task, Zhu et al. [12] proposed a set of connected component features used by a cascade of classifiers trained with Adaboost algorithm. In [7], the degree of "character-likeness" of each single-character-image is estimated using SVM and mesh feature describing local densities of black pixels. González and Bergasa [18] proposed a complete solution for reading text in natural images consisting of text location and text recognition stage. Letter candidates are extracted by combining MSER region detector and locally adaptive thresholding. Non-text objects are discarded based on geometric features, while character recognition is based on KNN classifier and gradient direction features. Mishra et al. [19] proposed method for scene text recognition that combines bottom-up and top-down cues. The first are extracted from character detections, while the second are obtained from language statistics. In [20], Wang et al. compared two approaches for scene text recognition: the first one using text detection and commercial OCR engine, and the second based on generic object recognition. It is demonstrated that latter solution has better performance. Our approach presented in this paper combines thresholding-based text extraction with CC (connected components)-based text detection which is exploited for selection of proper extraction hypothesis.

3. Choice of color space and color distance measure

3.1. IHLS color space

Choice of color space and color distance measure can strongly influence on success of text extraction step. Most algorithms are performed in RGB color space and color spaces derived with transformation from Cartesian coordinate system to polar (cylindrical) coordinate system (HSI, HSV, HSL etc.). Color representation in terms of hue, saturation and intensity better corresponds to the human intuition. Independence of these components enables separation of chromatic (hue and saturation) and achromatic (intensity) information. Perez and Koch [21] showed that hue is invariant to uniform scaling and uniform shifting in RGB space, while saturation is invariant to uniform scaling. Considering that uniform RGB scaling is caused by shadowing and uniform RGB shifting is caused by highlights, it is obvious that hue is robust to both artifacts while saturation is robust to shadowing. These properties justify transformation from RGB to HSI color space because it enables scene text extraction in images degraded with shadows and highlights. However, when working with HSI color space instability of components should be taken into account [22]: hue is unstable at low saturation and low intensity, while saturation is unstable for low intensity levels. Hanburry and Serra [9] discuss problems appearing in aforementioned usual variants of polar color spaces. Initial conic or biconic shape of these color spaces was expanded to cylindrical form in order to eliminate need for checking valid color coordinates, that is, coordinates lying in the color gamut. Cylindrical shaped space is achieved by defining saturation as ratio of distance from achromatic axis to the maximum distance for the corresponding intensity or lightness level. Normalization of saturation with intensity (lightness), caused by expansion to cylindrical form, generates two unwanted effects. The first one is a possibility that pixels with very low or very high intensity have high saturation levels what is contrary to the definition of color saturation. The second one is a dependence of saturation on intensity (lightness) that contradicts to the requirement for separating chromatic (hue and saturation) from intensity information. These disadvantages are avoided in IHLS color space by removing the normalization of saturation with lightness. In this way, saturation instability for low intensities is reduced. For this reason, IHLS color space is employed in this paper. Conversion from RGB to IHLS color space is given by [23]:

$$L = 0.2126R + 0.7152G + 0.0722B \quad (1)$$

$$S = max(R,G,B) - min(R,G,B) \quad (2)$$

$$H' = \arccos \frac{R - \frac{1}{2}G - \frac{1}{2}B}{(R^2 + G^2 + B^2 - RG - RB - GB)^{\frac{1}{2}}} \quad (3)$$

$$H' = \begin{cases} 360° - H', if\ (B > G) \\ H', \qquad otherwise. \end{cases} \quad (4)$$

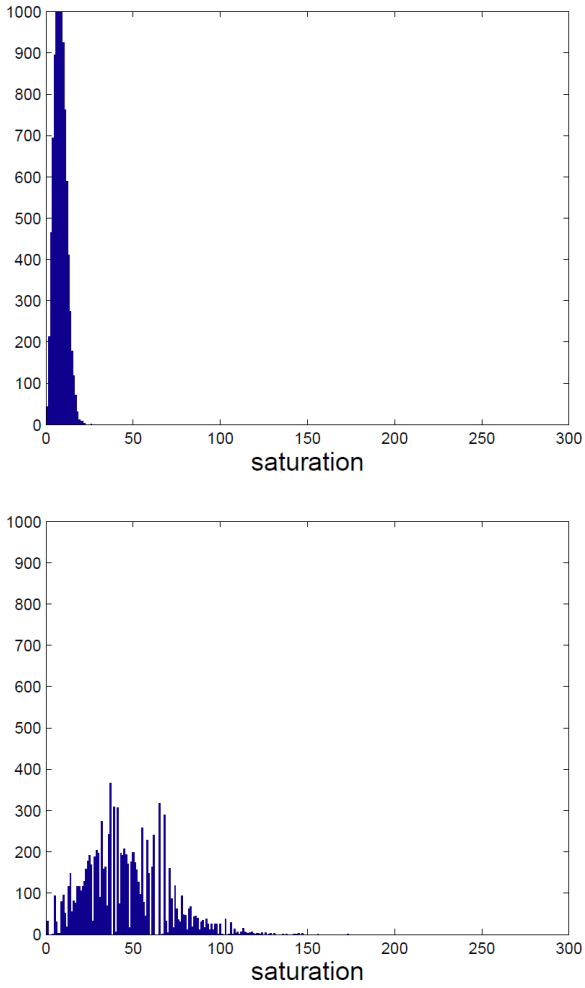

Figure 1. Saturation histogram for IHLS color space, L=10 (top) and HSI color space, I=10 (bottom)

To illustrate saturation stability, Fig. 1 shows histograms for saturation defined according to HIS and IHLS transformation. Histograms are calculated from 100x100 pixel samples obtained by adding additive white Gaussian noise to fixed RGB components. It is obvious that saturation component in IHLS shows significantly smaller degree of instability in comparison with saturation in HSI color space.

3.2. Cylindrical Color distance

The choice of color distance measure plays important role in ability to handle different degradations found in scene text images. Taking into account that colors are represented as 3D vectors it is possible to use measures such as Euclidean distance, Canberra distance, cosine distance etc. Mancass-Thillou and Gosselin [8] combine Euclidean distance and cosine-based similarity in RGB color space to exploit their complementarity: the first one handles intensity difference while the second reveals the change in chromaticity. In [24], Euclidean distance and several variations of angle distances are evaluated in RGB color space using k-means algorithm. Euclidean distance shows best performance in text extraction for RGB color space. However, this conclusion can't be generalized for polar color spaces because Euclidean distance doesn't correspond to their cylindrical nature. Cylindrical distance, proposed in [22], is more appropriate because it takes into account angular differences. For two pixels (H_i, S_i, L_i) and (H_j, S_j, L_j), cylindrical distance is defined as:

$$d_{cylindrical}(i,j) = \sqrt{d_{chroma}^2 + d_{ightness}^2} \quad (5)$$

$$d_{chroma}(i,j) = \sqrt{S_i^2 + S_j^2 - 2S_iS_j\cos\Theta} \quad (6)$$

$$d_{ightness}(i,j) = |L_i - L_j| \quad (7)$$

$$\Theta = \begin{cases} \Delta, & \text{if } (\Delta < 180°) \\ 360° - \Delta, & \text{otherwise} \end{cases} \quad (8)$$

$$\Delta = |H_i - H_j| \quad (9)$$

where $H \in [0°, 360°]$, $S \in [0, 255]$, $L \in [0, 255]$; d_{chroma} is the distance between two pixels in chromatic plane (chromatic distance), $d_{lightness}$ is lightness difference. Because of hue and saturation instability in case of HSI color space it is necessary to discriminate achromatic and chromatic pixels. For achromatic pixels, the intensity is only reliable component and cylindrical distance is reduced to intensity difference. In IHLS color space, saturation component is reliable, and although hue still has increased noise sensitivity for low saturation or lightness, it is not convenient to fully discard chromatic distance. Because of that, we didn't perform differencing of achromatic and chromatic pixels when calculating cylindrical distance in IHLS color space. In this paper, chromatic distance (6) is used for representing of chroma component: chromatic distance from background color is calculated for each pixel. After normalization the obtained values are thresholded using Otsu method giving the first text extraction hypothesis. The second extraction hypothesis is generated by thresholding of lightness component (7). Our choice is conditioned by the fact that the chromatic distance represents chromatic information, that is, hue and saturation components that are complementary to lightness information with respect to degradations in scene text images. Another reason is its suitability to polar color spaces.

4. The proposed method

The flowchart of the proposed text extraction method is shown in Fig. 2. The first step is generation of two text extraction hypotheses: one based on the chromatic distance, that is, chroma component, and the second based on the lightness component. In this step, Otsu thresholding is performed on both components in order to generate two binary images that represent candidates for final extraction results. After this step the hypothesis verification is performed using SVM classifier. For each CC, it is estimated the degree of "character–similarity". Segmentation with higher

average "character–similarity" value is chosen and sent to recognition stage done by OCR software. Details are described in next subsections.

4.1. Generation of text extraction hypothesis

In comparison with general image segmentation task, separating characters from background (text extraction step) requires more accuracy in order to avoid recognition errors. Due to variety of scene text images a promising strategy is a multi-hypotheses approach where two or more segmentations are performed and better result is chosen for recognition [8]. In our work, Otsu thresholding of chroma and lightness in IHLS color space is used for generation of two hypotheses. Lightness thresholding is appropriate for scene text images with even lightning and good contrast, while chroma thresholding, thanks to properties of saturation and hue component mentioned in Section 3.1, has potential to extract characters in presence of shadowing and highlights. Fig. 3 describes generation of text extraction hypothesis. In the first step, background color is estimated as color that most frequently appears on the edges of input image. The same approach is used in [25] resulting with high accuracy on ICDAR 2003 dataset. In order to include chroma component we first calculate chromatic distance of every pixel to background color. For image pixel on the i-th row and the j-th column with components $(H_{ij},\ S_{ij}, L_{ij})$ its chromatic distance from background color (H_{BG}, S_{BG}, L_{BG}) is given by:

$$chromatic_{distance(i,j)} = \sqrt{S_{ij^2} + S_{BG^2} + 2S_{ij}S_{BG}\cos\Theta}$$

$$\Theta = \begin{cases} \left|H_{ij} - H_{BG}\right|, \text{if } \left|H_{ij} - H_{BG}\right| < 180° \\ 360° - \left|H_{ij} - H_{BG}\right|, \quad \text{otherwise.} \end{cases} \qquad (10)$$

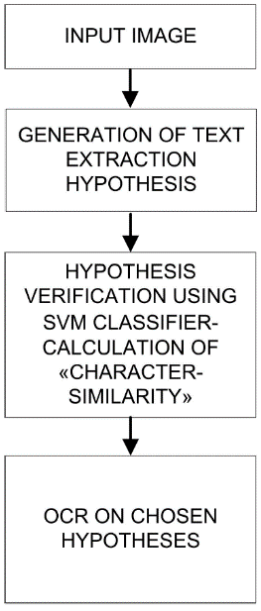

Figure 2. Stages of proposed text extraction method

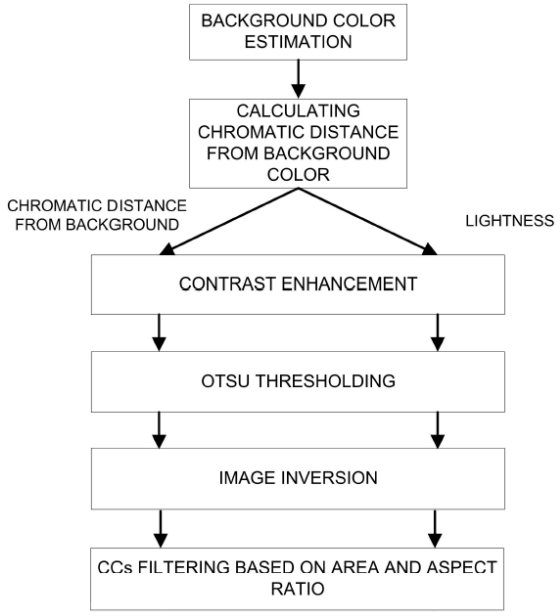

Figure 3. Flowchart of the algorithm for text hypothesis generation

Matrix $chromatic\_distance$ can be visualized as grayscale image: brighter areas correspond to pixels with greater chromatic distance from the background color. An idea is inspired by the fact that text is distinguishable from background which should be manifested as significant chromatic distance (difference between character color and background color). Similar approach is presented in [26] where text is extracted using pixel distance from text color that is selected using the camera focus. In our work, need for user hint is eliminated by calculating distances with respect to the background color. After contrast enhancement, image binarization is performed using well-known Otsu method on chromatic distance and lightness component. Although this segmentation method is rather simple, it performs quite well on ICDAR 2003 test set [25]. Instead of dealing with more complex segmentation methods, here we decide to move the focus to the generation of hypotheses that have potential for correct text extraction. It is assumed that one of the components should result with sufficient degree of histogram bimodality that enables successful segmentation with Otsu method. Image inversion is applied if there are white text pixels on black background. The aim is to obtain black characters on white background, mainly to improve recognition results with OCR software. The choice of inversion is based on the ratio of black and white pixels on image borders. At this stage CCs are filtered using a heuristic: CCs having too large or too small area and aspect ratio are rejected. In this way, obviously non-character CCs are excluded from further processing. Fig. 4 shows Otsu thresholding results obtained for lightness and chroma. In given example, it is obvious that chroma gives correct character extraction while lightness completely fails. Next steps perform selection of better extraction result using SVM based text detection.

Figure 4. Image segmentation: original image(top), chroma values, that is, pixel distances from the background color calculated with chromatic distance (middle left), lightness values(middle right), Otsu thresholding for chroma (bottom left) and lightness (bottom right)

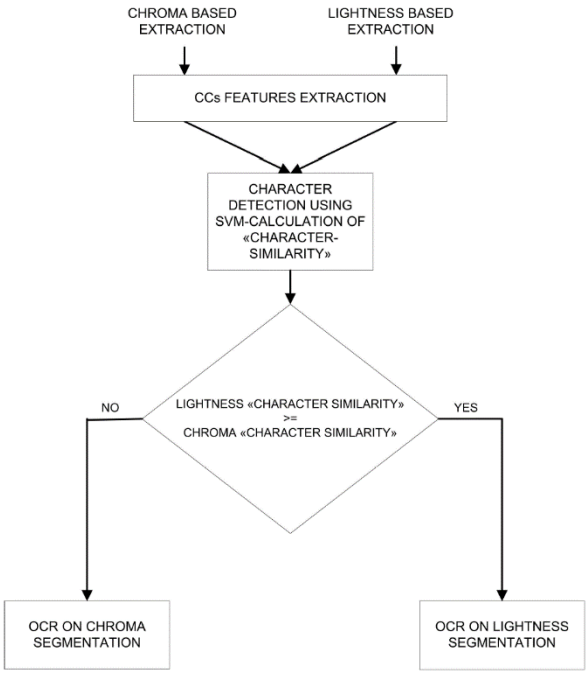

Figure 5. Hypothesis verification using SVM classifier

4.2. Hypothesis verification using SVM classifier and OCR on chosen hypothesis

After chroma and lightness thresholding the choice between two hypotheses should be done. A decision can be based on recognition step where segmentation resulting with higher recognition rate is taken as a solution [8]. This implies that for a given word we have to know ground truth to calculate the recognition rate. This approach also requires to perform recognition step on the multiple extraction results what increases the computational complexity. In this paper, character detection is used as a tool for verification of hypotheses generated in previous steps. Fig. 5 illustrates this procedure. Connected components are obtained as text extraction results for chroma and lightness component. Their geometrical features are extracted and used as input to support vector machine (SVM) classifier which gives degree of "character-similarity", that is probability of connected component to be a character. Extraction result with higher average "character-similarity" is chosen as valid result and sent to OCR processing stage. SVM classifier is widely used in literature dealing with text localization, extraction and recognition. There are several reasons for choosing SVM classifier that are stated in [27]. SVM has better generalization performance in comparison with other techniques like neural networks and radial basis function networks. It shows robustness in case of moderate training set. SVM works well in high-dimensional space and it is trained relatively easy.

The SVM classifier is based on statistical learning theory aiming at maximizing margin, that is, the distance between hyperplane separator and closest points of each class. If we take training set $D = \{x_i, y_i\}$, $x \in [0,1]$, $y \in -1$, 1 , $i \in [1,n]$ with $\boldsymbol{x_i}$ as input vector, and y_i as target label, linear SVM should satisfy the following condition:

$$y_i = (\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1 \quad (11)$$

where $\boldsymbol{w}$ is the weight vector, b is bias, and $\frac{2}{\|\boldsymbol{w}\|}$ is geometric margin. Finding maximum margin is a constrained optimization problem:

$$min(\frac{1}{2}\|\boldsymbol{w}\|^2) \text{ subject to } y_i = (\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1. \quad (12)$$

By introducing Lagrangian multipliers the problem is converted to:

$$\text{maximize } L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i}\boldsymbol{x_j}$$

subject to $0 \leq \alpha \leq C, i = 1, \ldots, n$ and

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \quad (13)$$

Coefficients (Lagrange multipliers) α_i are estimated from training points, while parameter C determines trade-off between size of margin and classification errors. Training data are often not linearly separable and this problem is solved by mapping of input features to higher –dimensional space using kernel function:

$$k(\boldsymbol{x}\boldsymbol{x_i}) = \Phi(\boldsymbol{x})\Phi(\boldsymbol{x_i}) \quad (14)$$

where $\Phi(\boldsymbol{x})$ is the feature mapping function. The most often used kernels are polynomial, radial basis function (RBF) and sigmoid. The decision function estimated by SVM is given by

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i k(\boldsymbol{x}\boldsymbol{x_i}) + b \quad (15)$$

The point $\boldsymbol{x}$ is assigned to one class or another according to the sign of $f(\boldsymbol{x})$. In this work, the value of $f(\boldsymbol{x})$ is used as degree of "character-similarity". The choice between two hypotheses is based on average "character-similarity" of extraction result:

$$average\_character\_similarity = \sum_{i=1}^{n} \frac{f(x_i)}{n} \quad (16)$$

where n is the number of connected components in extraction hypothesis, $\boldsymbol{x_i}$ denotes feature vector for the ith connected component and $f(\boldsymbol{x_i})$ is the value of the

decision function. Hypothesis resulting with higher $average\_character\_similarity$ is chosen for OCR processing. We used standard Support Vector Machine with Radial Basis Function (RBF) kernel:

$$k(\boldsymbol{xx_i}) = e^{-\gamma\|x-x_i\|}, \gamma = \frac{1}{2\sigma} \quad (17)$$

where σ is standard deviation. The values of parameters C (13) and γ are chosen using grid search and cross-validation procedure on the training set which consists of 925 non-characters and 1028 characters manually extracted from the training set of ICDAR 2011 word recognition task. Features used by classifier should enable good discrimination between connected components representing characters and non-characters. In this work, we select a set of 7 geometrical features (equations (18)-(24)) widely used in the literature ([6], [12], [28], [29], [30]) for text detection task. They are used as input in SVM classifier in order to calculate the degree of character similarity:

1. Aspect ratio is used for elimination of too long components:

$$AR(CC) = min\big(width(CC)/length(CC), length(CC)/width(CC)\big). \quad (18)$$

In this paper, the focus is put on the Latin alphabet where characters have aspect ratio in certain range of values. Arabian and Bangla text samples will be considered in future work.

2. Occupy ratio is defined as ratio of CC area to bounding box area:

$$OR(CC) = Area(CC)/Area(BoundingBox(CC)) \quad (19)$$

This feature removes CCs with too many or too few pixels in the bounding box.

3. Compactness enables elimination of non-character components with high complexity of contour shape:

$$Comp(CC) = Area(CC)/Perimeter(CC)^2. \quad (20)$$

4. Filled area is defined as:

$$FA(CC) = \frac{|Area(CC) - imfill(CC)|}{Area(CC)} \quad (21)$$

where $imfill(CC)$ is a function filling the holes in CC's. This feature is useful for character/non-character discrimination because text components usually don't contain large holes in comparison with their areas.

5. Surface area to convex hull ratio is defined as:

$$SC\_ratio = \frac{Area(CC)}{Area(ConvexHull(CC))}. \quad (22)$$

6. Skeleton length to perimeter ratio is used for text detection in [6]:

$$Skeleton\_LPR = \frac{length(skeleton(CC))}{Perimeter(CC)}. \quad (23)$$

7. Normalized stroke deviation is calculated as:

$$SWdev = \frac{std(Stroke\_Width(CC))}{mean(Stroke\_Width(CC))}. \quad (24)$$

Nearly constant stroke width is typical for characters allowing separation from other objects in the scene. The described features are calculated for each CC and their values are used as input to the SVM classifier which gives the degree of "character-similarity" for each CC. As it is shown in Fig. 5, the hypothesis yielding higher $average\_character\_similarity$ is chosen as text extraction result and processed with commercial OCR software. Text recognition results obtained in this way are used for performance comparison with other methods proposed in the literature.

5. Results and discussion

Evaluation of the proposed method was performed on the publicly available dataset from the ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images ([31], [32]). The test set (different from training set) of this database contains 1189 scene text images covering different kinds of degradations caused by imaging conditions (low resolution, blur, observation angle) and object or environment properties (complex background, uneven lightning, shadows, different fonts etc.). This dataset was used for the purpose of comparison with results presented by other authors on the same test set [31]. The performance evaluation on the test set of ICDAR 2013 Robust Reading Competition Challenge 2: Reading Text in Scene Images will be done in the future work. For evaluation of text extraction performance, we used two measures suggested in [31]. The first one is normalized edit distance between recognized and ground truth string where insertions, substitutions and deletions have equal cost. The second one is correct recognition rate representing percentage of correctly recognized words, that is, the number of cases where edit distance is equal to 0. After obtaining binarized image as text extraction result, OCR engine is employed to recognize characters and calculate previously mentioned performance measures. In this paper, we use commercial ABBYY Fine Reader 11 OCR software. Because of more objective evaluation of text extraction, we didn't use any dictionary or language models. Character set used for recognition consists of 103 characters including English letters, digits and 41 other characters (punctuation marks and special characters). In order to justify our approach we first extracted results obtained with Otsu thresholding of chroma and lightness component (Table 1). Extraction with lightness component yields obviously better performance in comparison with chroma.

Table 1. Text extraction results for chroma and lightness

Distance	Total Edit Distance	Correct Recognition(%)
Lightness	379.4	57.4
Chroma	830.4	25
Chroma+Lightness (ideal case)	315.7	62

Figure 6. Example where lightness outperforms chroma: original image(top), extraction based on chroma (middle), extraction based on lightness (bottom)

Figure 7. Example where chroma outperforms lightness: original image(top), extraction based on chroma (middle), extraction based on lightness (bottom)

Table 2. Complementarity between chroma and lightness according to edit distance: e_d_lightness is edit distance obtained for lightness and e_d_chroma is edit distance obtained for chroma

	Number of words
$e\_d\_lightness \leq e\_d\_chroma$	626 (54.2%)
$e\_d\_lightness \geq e\_d\_chroma$	104 (8.8%)

Although chroma thresholding exhibits poorer performance, it is interesting to analyze the degree of its complementarity with lightness regarding correct extraction. Further analysis (Table 2) reveals that in 626 (54.2%) images from test set lightness results with lower edit distance (Fig. 6). On the other hand, chromatic distance gives more extracted characters (Fig. 7) in 104 (8.8%) images. Similarly, with respect to correct recognition rate (Table 3) lightness extracts 439 (36.9%) words that are failed by chromatic distance, while reverse situation appears in 54 (4.5%) cases.

These results can be explained by the fact that chroma component, thanks to properties of hue and saturation component mentioned in Section 3, is robust to certain image degradations like shadows and highlights for which lightness fails. It is clear that combination of chroma and lightness segmentation has potential to improve the overall performance if proper binarization is chosen for given example. In ideal case, where for every image appropriate component is used, the total edit distance would be lowered from 379.4 to 315.7 (Table 1) and the correct recognition rate would be improved for approximately 4.5% in comparison with the approach using the lightness component only. Results of the proposed approach, where correct extraction is chosen using SVM classifier, are shown in Table 4. Compared to usage of lightness component, there is an improvement in total edit distance (decrease from 379.4 to 364.8) and correct recognition rate (from 57.4% to 59.7%). In comparison with results of other methods presented in [31] and [18], our algorithm clearly outperforms them in correct recognition rate.

Table 3. Complementarity between chroma and lightness according to correct recognition rate: extracted_lightness is set of words correctly recognized after lightness segmentation, while extracted_chroma is set of words correctly recognized after chroma segmentation

	Number of words
$extracted\_lightness \cap \overline{extracted\_chroma}$	439 (36.9%)
$extracted\_chroma \cap \overline{extracted\_lightness}$	54 (4.5%)

Table 4. Text extraction results of the proposed method in comparison with other approaches reported in [31]

Method	**Total Edit Distance**	**Correct recognition(%)**
Proposed method	364.8	59.7
González and Bergasa [18]	639.15	46.9
TH-OCR System	176.23	41.2
KAIST AIPR System	318.46	35.6
Neumann's method	429.75	33.11

Regarding total edit distance our algorithm is better than Neumann's method and method proposed by González and Bergasa (Table 4). It is noticed that character set affects the value of total edit distance: OCR engine often recognizes no character CCs as special characters and in that way the edit distance is increased. With character set reduced to letters and digits the number of such cases would be lowered. It is also interesting that correct recognition rate would be increased for 3.1% if case sensitivity is not taken into account. It should be noted that we haven't found publications, except overview [31], with more detailed description of TH-OCR and KAIST AIPR systems. Therefore we don't have insight in all details (dictionary usage, character set etc.) that are necessary for more comprehensive comparison with our method. Regarding text recognition stage, in Neumann method this step is performed using SVM classifier and language model. González and Bergasa employ KNN classifier with gradient features. Errors are later corrected using unigram language model. In THOCR and KAIST AIPR systems, commercial OCR programs are used (TH-OCR doesn't use any language model).

In the proposed method commercial OCR is used without any dictionary and language models. Histogram of normalized edit distances is shown on Fig. 8. Two peaks appear at the edit distance value equal to 0, corresponding to correctly recognized words, and value equal to 1 that represents cases where no character is recognized correctly. It is obvious that the proposed method has a tendency to work in binary fashion: words are either correctly extracted or totally missed with relatively small proportion of partially recognized examples. González and Bergasa [18] observe similar effect for their method on the same test set. In our case this can be explained by complementarity between chroma and lightness: if both of them fail to segment characters from background, most probably no character will be recognized resulting with edit distance equal to 1; otherwise, one of the distances usually generates proper extraction enabling recognition of the whole word. It should be noted that competing methods use segmentation algorithms (MSER, Niblack thresholding) that are computationally more expensive than Otsu thresholding used in our method. Despite that we obtained competitive results thanks to utilization of component complementarity and usage of SVM-based text detection for hypotheses verification. In this way the problem burden is distributed to text extraction and text detection steps.

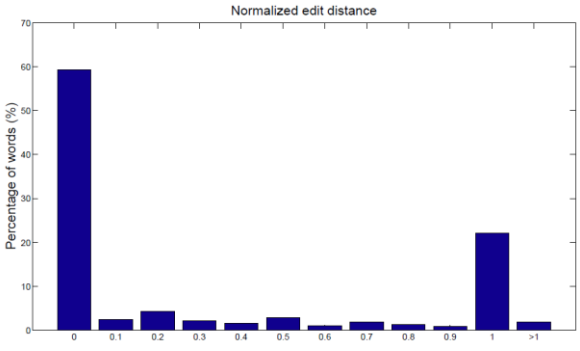

Figure 8. Histogram of normalized edit distances

6. Conclusion

In this paper we propose a scene text extraction method based on the multiple-hypotheses approach where SVM based character detection is used as a verification tool. Hypotheses are generated using Otsu thresholding of lightness and chroma in IHLS color space taking into account complementarity of these components regarding some kind of degradations. The choice of IHLS color space is motivated by its saturation definition that eliminates instability of this component at low lightness levels. Hypotheses verification is performed through estimation of "character-similarity" for each CC. For that purpose, SVM classifier with RBF kernel function is used. Experimental results confirmed complementarity assumption and demonstrated that our approach successfully exploits it giving performance improvement in comparison with usage of lightness component. Results are compared with recent competing methods and our algorithm showed best performance regarding correct recognition rate at the expense of higher total edit distance. It should be noted that these results are obtained using Otsu thresholding that is computationally less expensive than segmentation methods used in competing algorithms. In future work, the proposed method would be tested on ICDAR 2013 test set. It will also be investigated whether usage of other segmentation algorithms could improve results. It would be interesting to include text line features for SVM classification in order to choose correct extraction result.

References

[1] **V. Fragoso, S. Gauglitz, S. Zamora, J. Kleban, M. Turk.** An application of combinatorial optimization to statistical physics and circuit layout design. In: *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, 2011, pp. 497–502.

[2] **C. Thillou, S. Ferreira, B. Gosselin.** An Embedded Application for Degraded Text Recognition. *EURASIP Journal on Applied Signal Processing*, 2005, Vol. 13, pp. 2127-2135

[3] Google Goggles. www.google.com/mobile/goggles.

[4] **K. Jung, K. Kim, A. Jain.** Text Information Extraction in Images and Video: A Survey. *Pattern Recognition*, 2004, Vol. 37, No. 5, pp. 977–997

[5] **C. Mancas-Thillou, B. Gosselin.** Natural Scene Text Understanding. *Vision Systems: Segmentation and Pattern Recognition. Vienna, Austria: I-Tech Education and Publishing*, 2007, ch. 16, pp. 307–332.

[6] **L. Neumann, J. Matas.** A method for text localization and recognition in real-world images. In: *Proc. of the 10th Asian Conf. on Computer Vision*, 2010, pp. 770–783.

[7] **T. Wakahara, K. Kita.** Binarization of color character strings in scene images using K-means clustering and support vector machines. In: *2011 International Conference on Document Analysis and Recognition (ICDAR)*, 2011, pp. 274–278.

[8] **C. Mancas-Thillou, B. Gosselin.** Color text extraction with selective metric-based clustering. *Computer Vision and Image Understanding*, 2007, Vol. 107, No. 1-2, 97–107.

[9] **A. Hanbury, J. Serra.** Colour image analysis in 3D-polar coordinates. In: *Proceedings of DAGM*, 2003, pp. 124–131.

[10] **W. Niblack.** An introduction to image processing. *Prentice-Hall*, 1986.

[11] **Y. Pan, X. Hou, C. Liu.** A hybrid approach to detect and localize texts in natural scene images. *IEEE Trans. on Image Processing*, 2011, Vol. 20, No. 3, 800-813.

[12] **K.-H. Zhu, F.-H. Qi, R.-J. Jiang, L. Xu.** Automatic character detection and segmentation in natural scene images. *Journal of Zhejiang University SCIENCE A*, 2007, Vol. 8, No. 1, 63–71.

[13] **D. Karatzas, A. Antonacopoulos.** Colour text segmentation in web images based on human perception. *Image and Vision Computing*, 2007, Vol. 25 No. 5, 564–577.

[14] **R. Lienhart, A. Wernicke.** Localizing and segmenting text in images and videos. *IEEE Transactions on*

Circuits and Systems for Video Technology, 2002, Vol. 12, No. 4, 256–268.

[15] **H. Dujmić, M. Šarić, J. Radić.** Scene text extraction using modified cylindrical distance. In: *Proceedings of 12th WSEAS conference on Automation & Information*, 2011, pp. 213–218.

[16] **C. Garcia, X. Apostolidis.** Text detection and segmentation in complex color images. In: *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2000, pp. 2326–2330.

[17] **L. Zini, A. Destrero, F. Odone.** A classification architecture based on connected components for text detection in unconstrained environments. *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance AVSS '09*, 2009, pp. 176-181.

[18] **A. González, L. M. Bergasa.** A text reading algorithm for natural images. *Image and Vision Computing*, 2013, Vol. 31, 176–181.

[19] **A. Mishra, K. Alahari, C. V. Jawahar.** Top-Down and Bottom-Up Cues for Scene Text Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2687–2694.

[20] **K. Wang, B. Babenko, S. Belongie.** End-to-end scene text recognition. *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1457–1464.

[21] **F. Perez, C. Koch.** Toward color image segmentation in analog VLSI: algorithm and hardware. *International Journal of Computer Vision*, 1994, Vol. 12, No. 1, 17–42.

[22] **D. C. Tseng, C. M. Chang.** Color segmentation using perceptual attributes. In: *Proc. of the 11th Internat. Conf. on Pattern Recognition*, 1992, pp. 228–231.

[23] **A. Hanbury.** A 3D-polar coordinate colour representation well adapted to image analysis. In: *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, 2003, pp. 804–811.

[24] **C. Mancas-Thillou.** Natural Scene Text Understanding. *PhD Thesis, Faculté Polytechnique de Mons*, 2006.

[25] **C. Thillou, B. Gosselin.** Color binarization for complex camera-based images. *Electronic Imaging Conf. of the Int. Society for Optical Imaging*, 2004, pp. 301–308.

[26] **E. Kim, S. H. Lee, J. H. Kim.** Scene text extraction using Focus of Mobile Camera. In: *Proceedings of 10th International Conference on Document Analysis and Recognition*, 2009, pp. 166–170.

[27] **K. I. Kim, K. Jung, J. H. Kim.** Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, Vol. 25, No. 12, 1631–1639.

[28] **S. Karaoglu, B. Fernando, A. Trémeau.** A Novel Algorithm for Text Detection and Localization in Natural Scene Images. *2010 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2010, pp. 635–642.

[29] **H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, B. Girod.** Robust text detection in natural images with edge-enhanced Maximally Stable Ex-tremal Regions. *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2609–2612.

[30] **B. Epshtein, E. Ofek, Y. Wexler.** Detecting text in natural scenes with stroke width transform. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2693–2970.

[31] **A. Shahab, F. Shafait, A. Dengel.** ICDAR 2011 robust reading competition challenge 2: reading text in scene images. In: *Proc. 11th International Conference of Document Analysis and Recognition*, 2011, pp. 1491–1496.

[32] Robust Reading Competition Challenge 2: Reading Text in Scene Images. http://robustreading.opendfki.de/wiki/SceneText.

Received November, 2013.