Extraction of Scene Text in HSI Color Space using K-means Clustering with Chromatic and Intensity Distance

MATKO SARIC, MAJA STELLA, PETAR SOLIC
Faculty of electrical engineering, mechanical engineering and naval arhitecture
University of Split
R. Boskovica bb, Split
Croatia
msaric@fesb.hr, mstella@fesb.hr, psolic@fesb.hr

Abstract: Text extraction is important step that strongly influences on the final recognition performance. This task is especially challenging in case of scene text which is characterized with wide set of degradations like complex backgrounds, uneven illumination, viewing angle, etc. In this paper we evaluated text extraction based on K-means clustering in HSI color space with chromatic distance and intensity distance. Obtained results are analyzed with respect to their complementarity in order to show potential for performance improvement. Comparison is also made with approach using cylindrical distance in K-means clustering of HSI image.

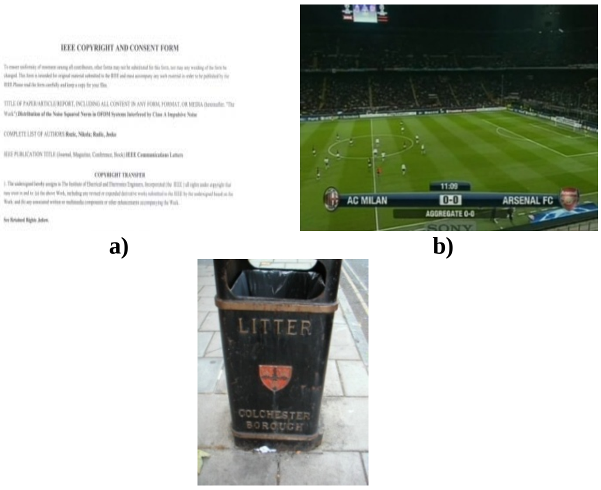
Key-Words: scene text extraction, K-means, chromatic distance, intensity distance

1 Introduction

Extraction of textual information from images and video enables range of different applications such as document analysis [1], automatic license plate recognition [2], sign detection and translation [3], content based image indexing [4], etc. In literature text is divided on next categories: text in documents, caption text and scene text [5]. Text in documents (Fig. 1a) has properties (high contrast, uniform font and background etc.) that enable easier character extraction and recognition. Caption text (Fig. 1b) refers to characters artificially added to image or video frame. Typical examples are subtitles or match results in sports video. Scene text (Fig. 1c) is integral part of recorded image or video frame, that is text found in everyday environment (for example label on the door or text on traffic signs). This type of text is characterized by variability of background, shape, font, etc.

Digital cameras and camera equipped smartphones give users opportunity to take photo or record video almost wherever and whenever they want. This also means that scene text present in our environment can be easily captured. Despite these advantages, problems arising from usage of these devices refer to sensor noise, viewing angle, blur, variable illumination etc. These conditions, in combination with fact that scene text doesn't have constraints like text in documents, make its extraction a challenging task.

In [5] text information extraction procedure is divided on 5 steps: detection, localization, tracking, extraction and enhancement, and recognition (OCR). In extraction step characters are segmented from background, that is, text pixels are separated from background pixels. Artifacts resulting from poor character extraction (parts of background, missing character parts etc.) can significantly lower accuracy of recognition performed by OCR software. Importance of extraction step is especially emphasized in case of scene text. Complex backgrounds, geometrical deformation, uneven illumination and other degradations make character extraction a demanding task that determines success of recognition stage.



a) b)

c)

Fig. 1 a) text in document b) caption text c) scene text

In [6] text extraction methods are divided in two categories: thresholding-based and grouping-based. Histogram thresholding [7], adaptive or local thresholding [8] and entropy based methods belong to first category. These methods have low computational requirements and successfully handle grayscale images or color channels separately, but they are not suitable in case of complex backgrounds and varying colors. Region-based, learning-based and clustering-based methods belong to second category. Region-based techniques includes methods like region-growing [9] and split-and-merge algorithm [10]. Their main advantage is inclusion of spatial information which is very important for character extraction, but parameter values have strong influence on performance. In learning-based methods text extraction is performed using well–known classifiers (multi-layer perceptrons, self-organizing maps). Main problem is creation of representative training database in order to encompass variability of scene text examples.

Clustering-based methods are based on assumption that pixels have tendency to form groups in chosen color space. K-means is one of the most often used algorithms for text extraction because of its speed and efficiency. In [11] scene text is extracted using K-means clustering in RGB color space with Euclidean distance and angle distance where authors choose better result based on feedback from recognition results. Number of clusters is set to 3 representing characters, background and character edges. Garcia and Apostolidis [12] exploited 4-means for text segmentation in HSV color space. In [13] authors propose text extraction method based on K-means clustering with modified cylindrical distance in HSI color space. Comparison is also made with K-means using cylindrical distance in HSI color space and Euclidean distance in RGB color space. Wakahara and Kita [14] used K-means clustering for generation of multiple extraction results where best result is chosen using SVM classifier.

In this paper we investigate text extraction performance of K-means algorithm in HSI color space with respect to two color distance measures: chromatic distance and intensity distance. The first one is distance between pixels in HS plane that reflects chroma difference, while the second represents absolute value of intensity difference. These measures are taken from the definition of cylindrical distance in HSI color space where overall pixel distance is calculated using chromatic and intensity distance. Cylindrical distance takes into account angular values and therefore it better corresponds to cylindrical nature of polar color spaces (HSV, HSI, HSL etc.) than Euclidean distance

In this paper scene text extraction performance is evaluated for both measures separately and comparison is made with approach using cylindrical distance. We also analyze degree of complementarity between results obtained with chromatic and intensity distances as possible direction to improve scene text extraction.

The rest of the paper is organized as follows. Section 2. shortly describes distance measures used in comparison. Section 3. discusses in more details K-means clustering. Results are presented in section 4 and conclusions are made in section 5.

2 Chromatic distance, intensity distance and cylindrical distance

Purpose of color distance measures is quantification of color difference. Choice of color distance has strong influence on text extraction performance. Representation of colors as 3D vectors enables usage of well-known distance measures for m-dimensional vectors as Manhattan distance or Euclidean distance

Cylindrical distance, introduced in [15], between two pixels in HSI color space with values (H_i, S_i, I_i) and (H_j, S_j, I_j) is defined as:

$$d_{cylindrical}(i,j) = \sqrt{d_{chroma}^2(i,j) + d_{intensity}^2(i,j)} \quad (1)$$

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

$$d_{intensity}(i,j) = |I_i - I_j| \quad (3)$$

$$\theta = \begin{cases} \Delta & if\ \Delta < 180^\circ \\ 360^\circ - \Delta & otherwise \end{cases} \quad (4)$$

$$\Delta = |H_i - H_j| \quad (5)$$

where $H \in [0^\circ, 360^\circ]$, $S \in [0,255]$, $I \in [0,255]$.

Value d_{chroma} (Fig.2a) refers to chromatic distance between two pixels, while $d_{intensity}$ refers to absolute value of intensity difference. Cylindrical distance $d_{cylindrical}$ (is represented as hypotenuse of right-angled triangle).

Instability of hue and saturation components is a problem that should be taken into account [15]:

- hue is meaningless when intensity is very low,
- hue is unstable when saturation is very low,

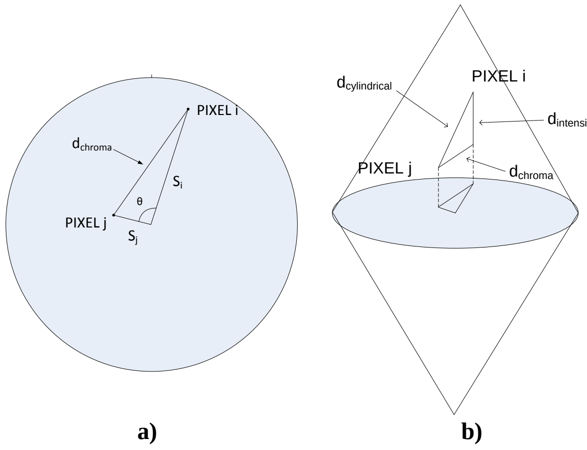


Fig. 2 a) chromatic distance b) cylindrical distance

- saturation is meaningless when intensity is very low.

Because of these properties it is necessary to perform distinction between chromatic and achromatic pixels. For chromatic pixels, that have stable hue and saturation, cylindrical distance is calculated according to equations (1)-(5). For achromatic pixels intensity is only relevant component. This implies that chromatic distance is unreliable and cylindrical distance is reduced to intensity difference $d_{intensity}$ (3).

Chromatic and achromatic pixels are usually classified by thresholding of saturation and intensity component. Achromatic pixels are defined with following conditions:

$$\begin{aligned} &intensity > 90\% \text{ max } intensity\,value, \\ &intensity < 10\% \text{ max } intensity\,value, \\ &saturation < 10\% \text{ max } saturation\,value. \end{aligned} \qquad (6)$$

Besides intensity difference, cylindrical distance also considers hue and saturation difference through chrominance information. These two components are robust to highlights and shadowing and hence they have potential to correctly extract characters in presence of such degradations. This is also the reason why HSI color space is chosen for scene text extraction in this paper.

3 Text extraction using K-means clustering

According to [6], clustering algorithms are considered as very efficient methods for scene text extraction. One of the most popular clustering techniques is K-means. Its main advantages are easy implementation and low computational requirements. K-means tries to minimize sum of distances between points and cluster centers that is represented by:

$$\sum_{j=1}^{k}\sum_{i\in S_j} distance\left(x_i^{(j)}, c_j\right) \qquad (7)$$

where *distance* is the chosen distance measure between point $x_i^{(j)}$ and the cluster centre c_j, S_j is set containing elements of cluster j and k is number of clusters. Algorithm consists of following steps:

1. In set of N points, corresponding to image pixels, choose k points as initial cluster centers (centroids) c_j

2. Assign each point to nearest cluster S_j based on its distance from cluster center c_j.

3. For each cluster S_j, compute a mean μ_j of each cluster and set the mean as new cluster center $\left(c_j = \mu_j\right)$

4. Repeat the steps 2 and 3 until the centroids no longer move

It should be noted that one of the main drawbacks of this algorithm is the need to fix the number of clusters. Regarding text extraction task 2 clusters (one representing character and second background) seems as logical choice, although in [11] authors used 3 clusters where third one corresponds to character edges.

In [16] author discussed a role to distance measure in K-means algorithm regarding text extraction. It is concluded that for RGB color space Euclidean distance gives best results, but angle distances can handle cases when Euclidean distance fails. Author also tested performance of K-means algorithm with Euclidean distance measure in different color spaces where RGB yields best results. It should be mentioned that Euclidean distance is not appropriate measure for cylindrical color spaces (HSI, HSV, HLS etc.).

4 Results

For evaluation purposes we used a test set from word recognition task in ICDAR 2011 Robust Reading Competition Challenge 2: Reading Text in Scene Images [17]. This database contains 1189 scene text images covering a broad set of problems like non-uniform backgrounds, different layouts, low contrast, variable illumination, low resolution etc. MATLAB was used for implementation of text extraction based on K-means algorithm with chromatic distance and intensity distance in HSI color space. The number of clusters is set to 2 (text and background). Binary images resulting from text extraction were processed with Google OCR engine Tesseract 2.04 in order to finally obtain recognized text.

As it is suggested in [17], the metrics for performance evaluation are edit (Levenshtein) distance and correct recognition rate. The first one is distance between ground truth string and string recognized using Tesseract, where deletions, substitutions and insertions have equal costs. Normalization is done by the number of characters in ground truth word. Second measure is percentage of correctly recognized words, that is, the number of words for which normalized edit distance is equal to zero.

Table 1 show results of K-means clustering for chromatic distance, intensity distance and cylindrical distance (taken from [13]) First column shows results reported in form of total edit distance calculated by summing normalized edit distance for each ground truth word. Second column presents correct recognition rate.

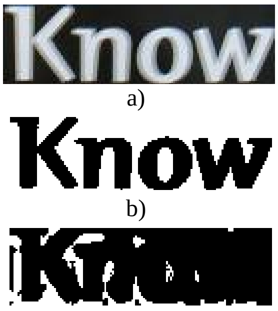
Distance measure	**Total Edit distance**	**Correct Recognition(%)**
Chromatic distance	1269,1	17.2
Intensity distance	669.5	47.2
Cylindrical distance [13]	764	40

Table 1 Scene text extraction results

It can be seen that intensity distance outperforms chromatic and cylindrical distance in both measures. This result reveals that intensity difference plays very important role in segmentation of characters from background. This observation is in accordance with conclusions presented in [16] and [13]: Euclidean distance, which mostly reflects changes in intensity component, performs best in scene text extraction task. Chromatic distance gives worst results what can be partially explained by instability of hue and saturation components. Cylindrical distance considers chroma and intensity difference, but this combination doesn't improve results in comparison with intensity distance only. Best result for this task on ICDAR 2011 database is reported in [18] where value of total edit distance is 639.15 and correct recognition rate is equal to 46.9%. Our results show that K-means clustering with intensity distance has similar performance.

	Number of words
e_d_intensity<e_d_chroma	670 (56.3%)
e_d_intensity>e_d_chroma	161 (13.5%)

Table 2. Complementarity between chromatic distance and intensity distance according to edit distance: e_d_intensity is edit distance obtained for intensity distance and e_d_chroma is edit distance obtained for chromatic distance

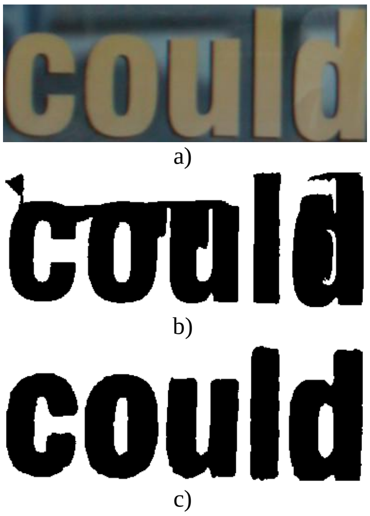


a)

b)

c)

Fig. 2 Example in which intensity distance gives better result than chromatic distance a) input image b) result of K-means with intensity distance c) result of K-means with chromatic distance



a)

b)

c)

Fig. 3 Example in which chromatic distance gives better result than intensity distance a) input image b) result of K-means with intensity distance c) result of K-means with chromatic distance

More detailed analysis (table 2.) shows that in 670 (56.3%) images from test set intensity distance results with better extraction (Fig. 3). Chromatic distance outperforms intensity distance in 161 (13.5%) images (Fig. 4). This indicates complementarity between these distances that can be exploited to further improve performance by combination of results.

Approach where examples that are not correctly extracted with intensity distance, would be segmented with chromatic distance could improve correct recognition rate for approximately 3% (from 47.2% to 50.5%). Despite overall efficiency of intensity distance, in cases when it fails chromatic distance is potential solution. This happens in images with certain kind of degradations like uneven illumination or shadows where robustness of hue and saturation plays important role.

5. Conclusion

In this paper we evaluated scene text extraction method based on K-means clustering in HSI color space with chromatic and intensity distance. Performances are compared with approach using cylindrical distance as metrics in K-means clustering. Interesting point is that intensity distance gives best performance outperforming cylindrical distance. This confirms conclusions previously presented in literature [16] that intensity or lightness is the component that most efficiently segments characters from background in case of scene text images. We also showed complementarity between results obtained with tested distances: although overall chromatic distance gives much worse results, it can successfully extract characters in some cases when intensity distance fails. This can be explained by fact that chromatic distance incorporates hue and saturation components that are robust to degradations like highlights and shadows. From presented results it can be concluded that fusion of results obtained with two or more distance measures in K-means algorithm has potential to improve scene text extraction performance.

References

[1] Y. Y. Tang, S. W. Lee, and C. Y. Suen, "Automatic document processing: a survey," *Pattern Recognition*, vol. 29, no. 12, p. 1931–1952.

[2] S. L. Chang, L. S. Chen, Y. C. Chung, and S. W. Chen, "Automatic license plate recognition," *IEEE Transactions on Intelligent Transport Systems*, vol. 5, no. 1, pp. 42-53, 2004.

[3] Y. Watanabe, K. Sono, K. Yokomizo, and Y. Okada, "Translation camera on mobile phone," in *Proceedings of International Conference on Multimedia and Expo*, 2003, pp. 177-180.

[4] M. Saric, H. Dujmic, V. Papic, N. Rozic, and J. Radic, "Player Number Recognition in Soccer Video using Internal Contours and Temporal Redundancy," in *Proceedings of the 10th WSEAS International Conference on Automation & Information (ICAI'09*, 2009, pp. 175-180.

[5] K. Jung, K. Kim, and A. Jain, "Text Information Extraction in Images and Video: A Survey," *Pattern Recognition*, vol. 37, no. 5, pp. 977-997, 2004.

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

[7] S. Messelodi and C. M. Modena, "Automatic identification and skew estimation of text lines in real scene images," *Pattern Recognition*, vol. 32, no. 5, p. 791–810, 1999.

[8] B. Gatos, I. Pratikakis, K. Kepene, and S. J. Perantonis, "Text detection in indoor/outdoor scene images," in *Proc. First Workshop of Camera-based Document Analysis and Recognition,*, 2005, p. 127–132.

[9] R. Lienhart and W. Effelsberg, "Automatic text segmentation and text recognition for video indexing," University of Mannheim, Technical Report, 1998.

[10] D. Karatzas and A. Antonacopoulos, "Colour text segmentation in web images based on human perception," *Image and Vision Computing*, vol. 25, no. 5, pp. 564-577, 2007.

[11] C. Mancas-Thillou and B. Gosselin, "Color text extraction with selective metric-based clustering," *Computer Vision and Image Understanding*, vol. 107, no. 1-2, pp. 97-107, 2007.

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

[13] M. Saric, H. Dujmic, and M. Russo, "Scene Text Extraction in HSI Color Space using K-means Algorithm and Modified Cylindrical Distance," *Przegląd elektrotechniczny*, vol. 89, no. 5, 2013.

[14] T. Wakahara and 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.

[15] D. C. Tseng and C. M. Chang, "Color segmentation using perceptual attributes," in *Proc. of the 11th Internat. Conf. on Pattern Recognition*, 1992, pp. 228-231.

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

[17] A. Shahab, F. Shafait, and 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.

[18] G. A and L. M. Bergasa, "A text reading algorithm for natural images," Image and Vision Computing, vol. 31, pp. 255-274, Mar. 2013.