TEXT EXTRACTION FROM RASTER MAPS USING COLOR SPACE QUANTIZATION

Sanaz Hadipour Abkenar and Alireza Ahmadyfard

Department of Electronic and Robatic Engineering,
Shahrood University of Technology, Shahrood, Iran

ABSTRACT

Maps convey valuable information by relating names to their positions. In this paper we present a new method for text extraction from raster maps using color space quantization. Previously, most researches in this field were focused on Latin texts and the results for Persian or Arabic texts were poor. In our proposed method we use a Mean-Shift algorithm with proper parameter adjustment and consequently, we apply color transformation to make the maps ready for K-Means algorithm which quantizes the colors in maps to six levels. By comparing to a threshold the text layer candidates are then limited to three. The best layer can afterwards be chosen by user. This method is independent of font size, direction and the color of the text and can find both Latin and Persian/Arabic texts in maps. Experimental results show a significant improvement in Persian text extraction.

KEYWORDS

Color space conversion, K-Means clustering, Mean-Shift algorithm, Quantization, Text extraction.

1. INTRODUCTION

Images are one of the most important media for transferring data. An image can have much higher impression than hundred lines of documents. Hence, image understanding and data extraction from images could be used for other tasks such as Machine learning [1].

Many organizations routinely use large sets of hard-copy graphic documents, including maps, engineering drawings, electrical schematics, and technical illustrations. Recent advances in computer technology allow graphic information to be stored and accessed more conveniently and cost-effectively in electronic form than on paper [2].

Maps are easily accessible compared to other geospatial data, such as vector data, satellite imagers, gazetteers, etc. Due to the availability of high quality scanners and existence of Internet, we can now obtain various maps in raster format for areas around the globe. By converting the text labels in a raster maps to machine editable texts, we can produce geospatial knowledge for understanding the region on map while its other geospatial data are not available. Moreover, a raster map can be registered to other geospatial data (e.g., imagers) and recognized texts from the map can be exploited for indexing and retrieval of the other geospatial data [3].

Text labels in raster maps link place names to their geographic locations. Texts in maps contain very important information, as converting the text labels in a raster map to machine editable text, helps produce geospatial knowledge for understanding a map region [4].

Jae-Kwang Lee et al. (Eds) : CCSEA, AIFU, DKMP, CLOUD, EMSA, SEA, SIPRO - 2017
pp. 77– 86, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70208

Finding a special place in a city map needs a strong map reader system. Therefore, improving existed map reader systems helps tourism industry. Today with the advent of auto-pilot cars, users prefer just to enter a name as an input to their cars GPS to be there. If we could extract texts from a map it could be useful in this case as well.

As stated earlier the most important use of text extraction is in GIS systems, but it could also be used in data mining, tourism and auto-pilot cars. Presenting a perfect map reader system which can improve precision (find the most texts with the least road sections) in minimum time and with minimum user interaction is the goal of almost all the researches in this field.

When we discuss about the streets of a city, color has an important role. In these kinds of raster maps, text, road lines, important building such as hospitals, schools, churches, mosques, etc. have been shown in different colors. Thus, color segmentation and quantization can help different layers separation and also text extraction.

In this paper we tried to improve the algorithm for extraction of Persian/Arabic and English text from geographical maps. Existence of points, subscripts, superscripts and some special parts of words which are in a lower or in a higher level from the words, discriminate Persian structure from English. Hence, existed methods on Latin texts are not applicable on Persian maps. Existing Persian methods work on grey scale maps. As most of the maps are colorful nowadays, they do not use color`s abilities. Moreover, their precision is low. In this paper we explain a method which can solve these issues. An algorithm is proposed in which we can extract text, specially Persian text, from colorful raster maps with the lowest error. We have tried to find the most possible words in the maps but least road lines and graphic symbols. This method is applicable for both English and Persian texts on the maps and is independent of font, size, direction and color of the texts.

The paper is organized as follows: In the next section we will review related works to text extraction. In the third section implementation of the algorithm will be described. Section four shows experimental results and last section shows conclusion of the proposed method.

2. RELATED WORKS

Fletcher and Kasturi described development and implementation of a new algorithm for automated text string separation which is relatively independent of changes in text font style and organizes individual characters. In their work, first connected components are produced. Then they use an area/ratio filter. Collinear component grouping and logical grouping of strings into words and phrases are respectively, their lateral steps in the proposed approach to separate text strings. The algorithm produces two images; One for texts and the other for graphics [5]. In their work Hough transform is used for character grouping and then text strings are extracted. As Hough transform only detect straight lines, their method cannot be applied to curved strings.

Chen and Wang presented a complete algorithm for extracting and recognizing numeral string on maps. Character extraction algorithm can segment slant and touching characters to their unique elements. Recognizing algorithm based on properties can also detect numeral characters with any size, position and direction. Discrimination property which is used here is simply detectable. In their proposed approach at first characters are extracted. After that recognition operation is applied. In this recognition holes, intended points, symmetric shapes and crossing points are recognized. Hough transform and a set of font and size dependent properties are also utilized for numeral strings detection [6]. However, this algorithm is not useful for alphabetic characters.

Velazquez and Levachkine proposed a method for separating and recognizing alphanumeric characters. In their method the map is segmented first, therefore all text strings, which contains touching symbols, strokes and characters, are extracted. Second, OCR-based recognition with artificial neural network (ANN) is applied to define coordinates, size and orientation of alphanumeric character strings in each case presented in map. Third, four straight lines or a number of curvatures which computed as a function of primarily recognized by ANN characters are extrapolated to separate those symbols that are attached. Finally, the separated characters are used as inputs into an ANN again to be finally identified [7].Velazquez and Levachkine's technique is presented for text detection in multi direction and curved strings. They divided their input documents into two equal columns. Each column is divided to some blocks based on connected components sizes to calculate linearity of local connected components and extract existing text strings.

In [8] Roy et. al. proposed a new approach for extracting unique text lines include pages of documents and also presented methods based on foreground and background textual character information. In their proposed approach, elements were recognized uniquely at first and were grouped to three clusters. Considering graph concept, the first three characters united to shape groups. Using background information between characters, the direction of added characters of a larger group is determined and based on these directions two candidate regions of different clusters formed. Finally, with the help of these candidate regions unique lines are extracted.

used oa caoting prec ies(b)1.8131(a)-7.8435(t)

Lee at. el. in [9] int31(e)3.74217(r,)-4.88632(4249.9-0.0579853(n7(I)8.598)-4.88491(e)-7.84357(x))-18(i)-0.05798p1(s)

(h)1.8131(e)3.74217(15(o)1.81(n)-364.043(s)6.6697214(g)1.8131(
araciogr(n)1.8131(u)-9.7726rrst4(g)13.3988(e)3.74217(r)-4.8863(

In4.88632( )-167.087(t)-0.0579853(h)-9.77357(e-4.88632( )-132.3

3. IMPLEMENTATION OF THE ALGORITHM

Figure (1) shows a part of Tehran map that we used in this work to report our results. The steps used to obtain the results are described in the following sub-sections.

Figure 1. A part of Tehran's map obtained from map.ketabeavval.ir

3.1. Mean-Shift algorithm

Initially, we applied Mean-Shift algorithm to reduce noise and to smooth the color of each region on the map. Mean-Shift algorithm considers the relation between colors in an image (pixel's position) and also considers the color of each pixel. It tries to change a cluster pixels' color to the mean of that cluster. As HSI color space provides a proper human understanding, we used this color space in our method. P(x, y) is the coordinates of pixel P in the image and H, S and I are the color of P.

To reduce noise in a map Mean-shift algorithm starts calculating the mean node for pixel P from N^{th} neighboring node $M(x_m, y_m, h_m, s_m, i_m)$. The position of the mean node contains mean value on each of the x, y, h, s and I axis of the N^{th} neighboring node to a local region. As explained in [14] if the distance between M and N become greater than a small threshold, Mean-Shift move N to M and calculate the mean node in the local area again. After convergence, the Mean-shift algorithm considers the values of h, s and i as the color of P(x, y) pixel. The result of applying mean-shift to Figure (1) is illustrated in Figure (2).

Figure 2. Applying Mean-Shift algorithm to the map shown in Figure 1.

3.2. Changing color space to Lab

Lab color space is a colorproof space with dimension of L for light, and a and b for colorproof dimensions based on non-linear compression coordinates. This color space contains all perceptible colors. This means that its expanse is larger than RGB and CMYK. One of the most important properties of Lab is that it is device independent and this senses that all of the colors are defined depend to their producing natures and the device in which they are shown. Figure (3) demonstrates changing color space process from RGB to Lab.

Figure 3. Changing color space from RGB to Lab for perception of colors and preparing it for K-Means.

3.3. K-Means algorithm

K-Means algorithm considers some points haphazardly as the mean points. Then with a distance measurement (usually it is Euclidean measure) each node distance to the nearest chosen node is calculated as the mean point and the new centre of gravity is found. The process is repeated again and again until mean points converge. The purpose of this algorithm is that i observations are classified in to k categories. We applied K-Means algorithm to generate an image that has a maximum of K colors. K-Means algorithm significantly reduces the number of colors in a map with maximizing the variance between classes.

Figure (4) indicates the results of applying K-Means algorithm on figure (3) which was obtained from color space transformation.

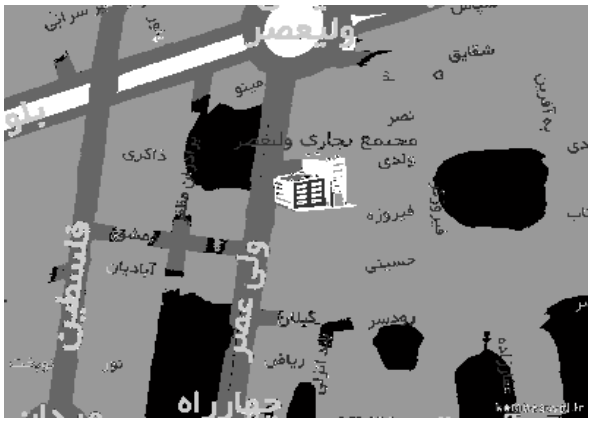

Figure 4. Applying K-Means algorithm on figure (3).

As can be seen from figure (4) all of the colors in the image are converted to K class and each of the regions is labeled to one of these K class. In our experiments K=6 was a good number for K and it had good results on other maps as well.

3.4. Calculating each region's area

As mentioned in previous section, a label is assigned to each region on the map. We consider each region which has a special label as a layer of the map and calculate the area of each region. Then we sort the areas in ascending format and find the median of the area's sizes. With the assumption that if the size of an area is smaller than the median it contains text part, the system chooses three layers as the possible text layers. In the last step the user choses the best layer from these three layers. This step needs user interference because K-Means is a supervised technique and its results may be different in each repetition. In other word, each layer may show different labels in each repetition.

3.5. Removing the largest connected component

In most of the maps, a long road line or a large non-text element is still in our best layer. We used the connected components analysis to find the largest connected component and remove this possible non-text region. System will ask the user whether or not to remove the connected component. If the user chooses yes, it will be omitted and if the user choose no it will be remained on the map. Figure (5) shows the results of choosing best layer between the possible layers.

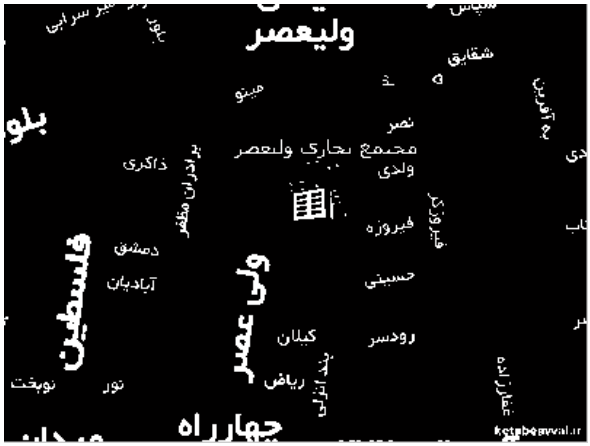

Figure 5. The best textual layer

4. EXPERIMENTAL RESULTS

We have done our experiments on 40 maps from Google, Yahoo, Tehranmaps, Ketabeavval, a number of maps from Tehran municipality's site and some other maps from different sources. Figures (6-10) show another example in which all steps of our algorithm are applied to an English map. This shows that the method is applicable to maps with Latin texts as well.

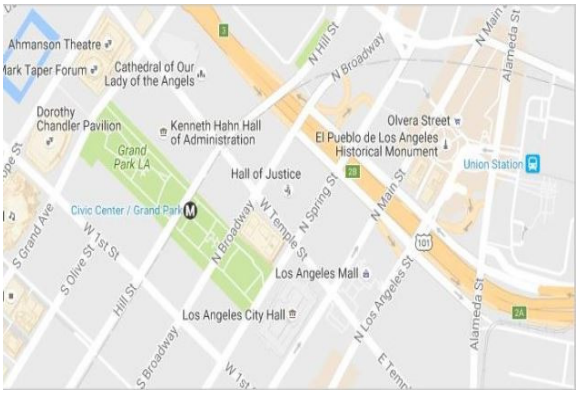

Figure 6. A map with English text obtained from maps.google.com

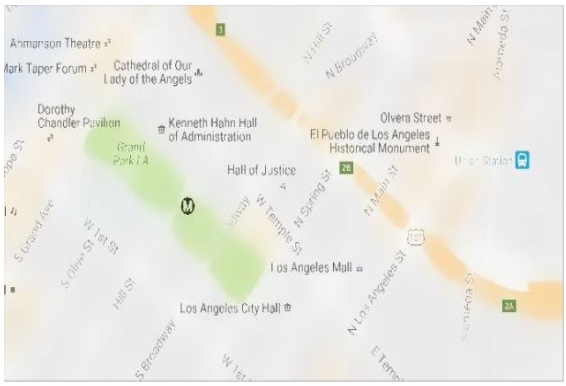

Figure 7. Applying Mean-Shift algorithm to map in figure (6).

Figure 8. Map of figure (7) after color transformation.

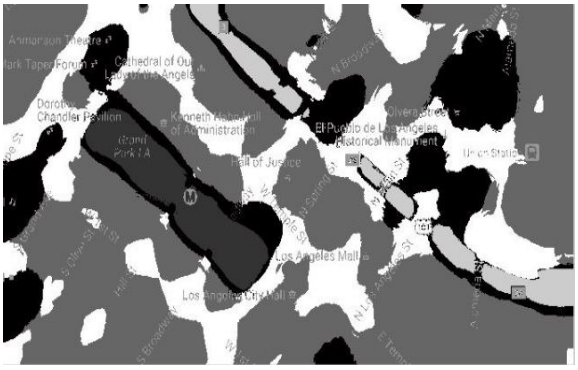

Figure 9. Results after applying K-Means algorithm

Figure 10. The best textual layer

As researches on Persian texts are very limited and the results of the previous works were poor, it was not possible to make a comparison. However, we calculated precision and recall percentages with the following formulas:

$$\text{Precision} = TP/(TP+FP) \quad (1)$$

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

In these formulas, TP (True Positive) refers to text areas which our system could recognize as text correctly. FP (False Positive) refers to non-text areas which our system detected them as text wrongly and FN (False Negative) refers to text areas which our system could not recognize them, wrongly. We counted each word as a candidate for text areas and considered each road line or each graphical symbol as a non-text area. Experimental results show %95.06 precision with %85.72 Recall. This confirms the applicability of the proposed approach for maps with Persian text with good precision. It should be noted that the proposed method is not limited to Persian/Arabic texts and can also work on maps with Latin text.

We are currently working on our approach to make the last step, which can find the best textual layer, automatic and make the texts ready for OCR software.

5. CONCLUSION

This paper presents an algorithm for extracting both Persian and English texts from raster maps. The proposed method is independent of font, size, color and direction of texts. We use a

Mean-Shift algorithm with proper parameter adjustment and apply color transformation to make the maps ready for K-Means algorithm, which quantizes the colors in maps to six levels. By comparing to a defined threshold the text layer candidates are then limited to three. The best layer is finally selected by user. Experimental results show %95.06 precision with %85.72 Recall for the proposed approach.

REFERENCES

[1] M. Tabassum and M. Shorif Uddin, (2011) "Extraction of ROI in Geographical Map Image," Journal of Emerging Trends in Computing and Information Sciences, Vol 2, No. 5, pp. 237-242.

[2] G. K. Myers and P. G. Mulgaonkar, (1996) "Verification-Based Approach for automated Text and Feature Extraction from Raster-Scanned Maps", Springer, Vol. 1072, pp. 190-203.

[3] Y. Y. Chiang and C.A. knoblock, (2010) ,"An approach for recognizing text labels in raster maps," International Conference on pattern recognition, pp. 3199-3202.

[4] Y-Y. Chiang and C.A. Knoblock, (2014) " Recognizing text in raster maps", GeoInfomatica, Vol 19, Issue 1, pp. 1-27.

[5] LA. Fletcher and R. Kasturi, (1988) "A robust algorithm for text string separation from mixed text/graphics images". IEEE Trans. Pattern Analysis and Machine Intelligence, Vol 10, Issue 6, pp. 910-918.

[6] L-H. Chen, J-Y. Wang, (1997) "A system for extracting and recognizing numeral strings on maps", Proceedings of the 4th international conference on document analysis and recognition, Vol 1, pp. 337–341.

[7] Vel´ azquez A, Levachkine S, (2004) "Text/graphics separation and recognition in raster-scanned color cartographic maps", Graphics recognition. Recent Advances and perspective, Springer, Vol 3088, pp. 63-74.

[8] Roy PP, LIados J, Pal U, (2007) "Text/graphics separation in maps", International Conference on computing Theory and Application, pp. 545-551.

[9] L.Li, G. Nagy, A.Samal, SC.Seth and Y.Xu, (2000) "Integrated text and line-art extraction from a topographic map", IJDAR, Vol 2, Issue 4, pp. 177-185.

[10] J . Pouderoux, JC . Gonzato, A . Pereira and P. Guitton, (2007) "Toponym recognition in scanned color topographic maps", 9th international conference on document analysis and recognition, Vol 1, pp. 531–535.

[11] PP. Roy, U. Pal, J. Llad´ os and F. Kimura, (2008) "Multi-oriented English text line extraction using background and foreground information", The eighth IAPR international workshop on document analysis systems, pp. 315–322.

[12] A. Kabir, A. Ghaffari, K. Kangarloo (2010), "Sepration of Persian text from scanned metropolitan maps", 17th Iranain Conference on Image processing and machine learning, pp 1-4.

[13] A. Kabir, A. Ghaffari, K. Kangarloo (2011), "A method based on distance conversion for revealing text in metropolitan map images", 20th Iranian Conference on Electrical Engineering, pp. 1-4.

[14] Y-Y. Chiang and CA. Knoblock, (2011) "A general approach for extracting road vector data from raster maps," IJDAR, Vol. 16. pp. 55-8.

AUTHORS

Sanaz Hadipour Abkenar studied electronic engineering in Guilan university (2011) for her bachelor degree and is now a Master student of communication Engineering in Shahrood University of Technology. Her main interest is digital image processing.

Alireza Ahmadyfard received Ph.D. in image processing and Computer vision from CVSSP (Center for Vision Speech and Signal Processing) at University of Surrey in 2002. He is director of Electrical Engineering Department in Shahrood university of technology. His research interests are digital signal processing, object recognition, image based inspection and human identification using biometrics.