Application of Change-Point Detection in Image Retrieval

Dongwei Wei
and Yuehua Wu

Department of Mathematics and Statistics
York University,
Toronto, Ontario, Canada
Email: weidw@mathstat.yorku.ca, wuyh@mathstat.yorku.ca

Xiaoping Shi

Department of Mathematics,
Statistics and Computer Science
St. Francis University,
Antigonish, Nova Scotia, Canada
Email: shermanshixp@gmail.com

***Abstract*—In this paper, we approach the image retrieval problem from a different angle by converting the problem into a change-point detection problem. An algorithm is introduced for detecting multiple change-points in distributions of a sequence of independently distributed random variables. By using this algorithm, a procedure is given for image retrieval. The proposed method is evaluated via two examples, which show that the proposed method for the image retrieval has satisfactory performance in terms of the quality of the retrieved images.**

Keywords–Change-point detection; Image retrieval; Contamination; Outliers

I. INTRODUCTION

The term "change-point" refers to a time moment or a spatial location at which the data generation process undergoes an abrupt change so that a different model needs to be used to characterize the generation mechanism after the change by [13]. Statistical studies of change-point problems started with [12] and have flourished especially since the 1980s (see the books [2] and [1] among others). Results of these studies have found applications in a wide range of areas such as quality control, finance, environmetrics, medicine, geographics, and engineering. Statistical models in change-point problems may vary in different application areas. The one used in this paper is the change-point in probability distribution, referring to a generic change of the distribution of observations before and after the change-point. Another popular one is the change-point in linear regression which includes the change-point in mean as its special case. As commented in [13], the essential difference between the model with change-points and the usual piecewise model is that the points of change in the latter are specified while in the former they are unknown and need to be estimated. In addition, for general change-point models, it is unknown whether change-points even exist, and when they exist, how many there are. This uncertainty adds to the difficulty and complexity in analyzing change-point models. Therefore detecting all change-points in a data series has become an important task in the analysis of change-points. It is well known that if there exist change-points, it is not appropriate to make a statistical analysis without considering their existence and the results derived from such an analysis may be misleading.

In digital image analysis, each grey-scale image consists of a number of pixels that are elements of a matrix (named as image matrix hereafter) (see [7] and [10] among others). Dimensions of the image matrix for a fine image are usually very large. If an image is corrupted, and hence the corresponding image matrix is no longer the one for the original image, the challenging problem is how to retrieve the original image or equivalently recover its image matrix. There is a rich literature on image retrieval which includes [3] [7] [8] [11] [14] among others. In this paper, we approach the image retrieval problem from a different angle. By considering each row and each column of the image matrix of this corrupted image as data sequences, we can convert the image retrieval problem into a multiple change-point detection problem. Thus the image retrieval problem may be solved by employing a multiple change-point detection method.

The paper is arranged as follows. In Section 2, we briefly introduce the problem of multiple change-points in distributions and then present the multiple change-point detection algorithm proposed in [16]. In Section 3, we give a procedure for image retrieval. In Section 4, we evaluate the performance of the proposed method via two examples. We complete this paper with some concluding remarks in Section 5.

II. MULTIPLE CHANGE-POINT DETECTION

Let $X_1, \cdots, X_n$ be independently distributed random variables with distributions $F_i, i = 1, 2, \cdots, n$, respectively. The problem of multiple change-points in distributions is that there exist $1 < k_1 < k_2 < \cdots < k_p < n$ such that $F_1 = \cdots = F_{k_1} \neq F_{k_1+1} = \cdots = F_{k_2} \neq F_{k_2+1} = \cdots = F_{k_3} \neq \cdots \neq F_{k_{p-1}+1} = \cdots = F_{k_p} \neq F_{k_p+1} = \cdots = F_n$. The task of multiple change-point detection is to find the number of change-points, i.e., to find p, and to determine the locations of these change-points, i.e., to estimate $k_1 < k_2 < \cdots < k_p$. Some work on detecting a change in distribution include [5], [6], [4], and [16] among others. The following material is mainly based on [16].

First, consider the single change-point detection problem. One needs to find out if there exists a change in distribution at k^* such that $F_1 = \cdots = F_{k^*} \neq F_{k^*+1} = \cdots = F_n$ with $1 < k^* < n$. It is noted that k^* is unknown. If $k^* = 1$ or n, we consider that there is no change in distribution.

Let

$$C_k(t) = \sqrt{\frac{k(n-k)}{n}} \left(\frac{1}{k} \sum_{i=1}^{k} \cos(tX_i) - \frac{1}{n-k} \sum_{i=k+1}^{n} \cos(tX_i) \right), \quad (1)$$

which is based on the real part of empirical characteristic function combining with the traditional cumulative sum method. If

 ISBN: 978-1-61208-358-2

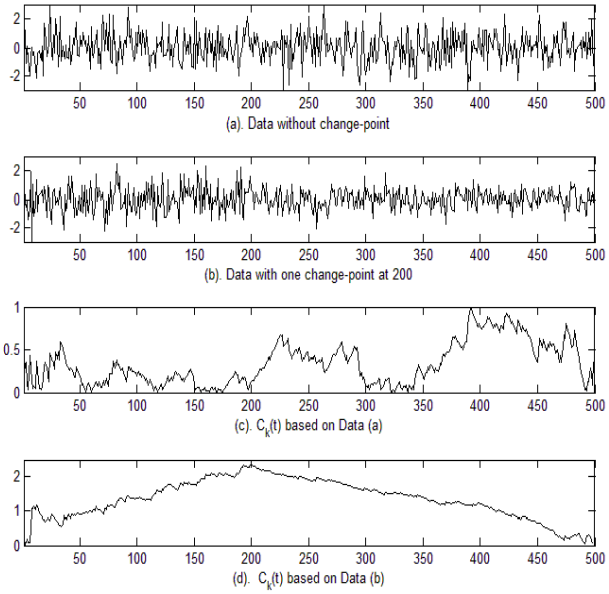

Figure 1. Top two panels: (a) data without any change in distribution; (b) data with one change in distribution at $k^* = 200$. Bottom two panels: plots of $C_k(t)$ with $t = 0.2$ based on the data displayed respectively in (a) and (b).

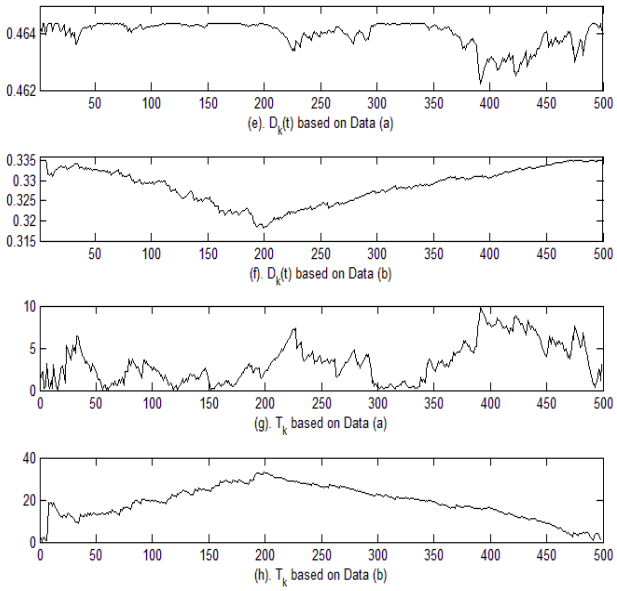

Figure 2. Top two panels: plots of $D_k(t)$ with $t = 0.2$ based on the data displayed respectively in (a) and (b) of Figure 1. Bottom two panels: plots of T_k based on the data displayed respectively in (a) and (b) of Figure 1.

k is the true change-point, i.e., $k = k^*$, the value of $|C_k(t)|$ is expected to be the largest.

Define

$$T_k = \int_t \omega(t) \frac{C_k(t)}{D_k(t)} dt, \tag{2}$$

where $\omega(t) = \frac{2}{\pi}\sqrt{(1-t^2)}$, $t \in [-1, 1]$,

$$D_k^2(t) = \frac{1}{n}\Big\{ \sum_{i=1}^{k} \Big(\cos(tX_i) - \frac{1}{k}\sum_{j=1}^{k} \cos(tX_j) \Big)^2 + \sum_{i=k+1}^{n} \Big(\cos(tX_i) - \frac{1}{n-k}\sum_{j=k+1}^{n} \cos(tX_j) \Big)^2 \Big\} \tag{3}$$

Illustrations of the statistics $C_k(t)$, $D_k(t)$ and T_k are plotted in Figures 1-2, where the true change-point is located at $k^* = 200$ and the sample size $n = 500$. For the case that there is no change-point, $F_1 = \cdots = F_n = N(0, 1)$, while for the case that there is a change-point, $F_1 = \cdots = F_k = N(0, 1)$ and $F_{k+1} = \cdots = F_n = N(0, 0.36)$.

It is shown in [16] that if $X_1, \cdots X_n$ are independent identically distributed random variables, under the assumption that there is no change in distribution, then

$$\lim_{n\to\infty} P\{A(\log n) \max_{1\le k\le n} T_k \le u + D(\log n)\} = \exp(-2e^{-u}) \tag{4}$$

where $A(x) = (2\log x)^{1/2}$ and $D(x) = 2\log x + 0.5\log\log x - 0.5\log\pi$.

To perform change-point detection, we first calculate the statistic $\max_k T_k$ and then compare it with the corresponding critical value which is computed by using (4). If $\max_k T_k$ is smaller than the critical value, no change-point is claimed, otherwise there exists a change-point which is estimated by $\hat{k} = \arg\max_{1<k<n} |T_k|$.

Now consider the multiple change-point detection problem introduced in the beginning of this section. In light of the iterated cumulative sums of squares algorithm in [9], [16] proposed an efficient and fast algorithm for detecting multiple change-points in distributions, which is given below. This algorithm will be used in our image retrieval procedure given in Section 4.

Let $X[l_1 : l_2]$ represent the segment $X_{l_1}, X_{l_1+1} \cdots, X_{l_2}$ with $l_1 < l_2$. Denote $T_k(t)$ that is computed in terms of $X[l_1 : l_2]$ by $T_k(X[l_1 : l_2])$. Define $k^*(X[l_1 : l_2])$ to be the point at which $M(X[l_1 : l_2]) \equiv \max_k T_k(X[l_1 : l_2])$ is attained. Let $CV(X[l_1 : l_2])$ be the critical value computed via (4). Denote the set of detected change-points by CP. The pseudo-code of the algorithm is as follows:

1) Set $l_1 = 1$ and $l_2 = n$. Calculate $M(X[l_1 : l_2])$ and $k^*(X[l_1 : l_2])$;
2) While $M(X[l_1 : l_2]) > CV(X[l_1 : l_2])$, repeat 3) − 12);
3) Set $k_{first} = k_{last} = k^*(X[l_1 : l_2])$;
4) Set $M_1 = M(X[l_1 : k_{first}])$ and $k_1 = k^*(X[l_1 : k_{first}])$;
5) While $M_1 > CV(X[l_1 : k_{first}])$, repeat 6);
6) $k_{first} = k_1; M_1 = M(X[l_1 : k_{first}])$;
7) Then set $M_2 = M(X[k_{last} : l_2])$ and $k_2 = k^*(X[k_{last} : l_2])$;
8) While($M_2 > CV(X[k_{last} : l_2])$), repeat 9);
9) $k_{last} = k_2; M_2 = M(X[k_{last} : l_2])$;
10) If $k_{first} == k_{last}$, there is only one change-point in $[l_1 : l_2]$, add it in CP and end the loop;
11) Else add the two candidate change-points in CP and then continue;

12) Reset $l_1 = k_{first}, l_2 = k_{last}$;

13) Sort change-point in CP. Denote the size of CP by N. Add 0 and n in CP. Then $CP = \{\ell_0 = 0, \ell_1, \ldots, \ell_N, \ell_{N+1} = n\}$ and $\ell_i < \ell_j$ if $i < j$.
14) For $j = 1, \ldots, N$, check whether a possible change-point exists between $[\ell_{j-1} + 1 : \ell_{j+1}]$, if yes, keep the jth change-point; if not, eliminate the jth change-point from CP.
15) Repeat $14)$ until CP does not change.

We name the above algorithm as *Algorithm WSW*. Here "WSW" is the acronym of the last names of the three authors of [16]. By applying this algorithm, one can not only estimate the number of multiple change-points (if the number is 0, there is no change in distribution) but also estimate multiple change-point locations.

III. IMAGE RETRIEVAL

First let us consider a noise contaminated black white image with scratches. If we treat each row (column) of the image matrix of this corrupted image as a data series, it is easy to see that each data series contains change-points that reflect the color changes. We remark that the locations of scratches do not correspond to locations of change-points, instead, they correspond to the locations of outliers. Thus a robust statistical change-point detection method may be applied to locate the true change-points, and hence the image matrix for the original image can be recovered. Since there are large number of rows even for a small image, it is important to employ a fast and efficient change-point detection method to tackle such a problem. *Algorithm WSW* (see the previous section) is suitable for this task.

For each row (column) of the image matrix, we employ *Algorithm WSW* to find out whether or not there exist multiple change-points and then estimate their locations if multiple change-points do exist. These multiple change-points are used to segment the data sequence in each row (column), which is the key in our image retrieval method. We are now ready to propose a procedure for image retrieval, which consists of the following steps:

1) Convert the noise contaminated black white image into the image matrix, denoted by A.
2) Use *Algorithm WSW* to detect multiple change-points for each row of A that divide the data sequence in each row into segments.
3) For each row of A, replace all the numbers within each segment by the segment median. Denote the resulting matrix by $A_r = (a_{ij}^{(r)})$.
4) Repeat steps 2) and 3) for each column of the matrix A. Denote the resulting matrix by $A_c = (a_{ij}^{(c)})$.
5) Generate a new matrix $B = (b_{ij})$ by combining A_r and A_c such that $b_{ij} = \min\{a_{ij}^{(r)}, a_{ij}^{(c)}\}$.
6) Repeat steps $2) - 5)$ to refine the matrix B. The final matrix is the restored image matrix. (Optional step)

We name this procedure as *Procedure CP*. Here "CP" stands for initials of "change-point".

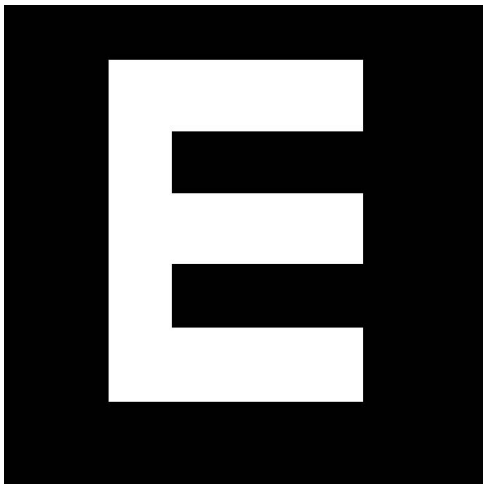

Figure 3. The image of the letter 'E'.

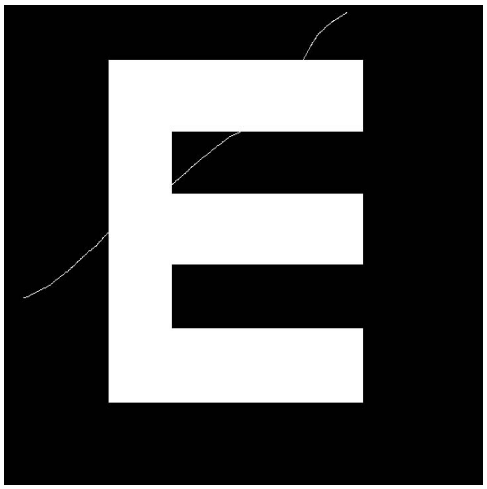

Figure 4. The image of the letter 'E' with some scratches.

IV. TWO EXAMPLES

In this section, we evaluate *Procedure CP* via two examples. We first focus on the example given in [15]. Consider Figures 3-5 below. The image of the letter 'E' is shown in Figure 3. We then add a scratche to the image of the letter 'E'. The resulting image is displayed in Figure 4. In order to examine if our procedure can restore the original image of the letter 'E' under even worse conditions, we contaminate the image displayed in Figure 4 by adding some noise. The resulting image is shown in Figure 5.

To proceed, we first convert the noised image of the letter 'E' with some scratches to the image matrix of dimensions 542×719. We then apply *Procedure CP* to this image matrix for image retrieval. The restored image is displayed in Figure 6, which shows that the image of the letter 'E' is retrieved successfully.

As a second example, we construct an image displayed in Figure 7, which apparently is more complex than the image of the letter 'E' as it is the combination of circle and triangle. Similar to the previous example, we add some scratches to

 ISBN: 978-1-61208-358-2

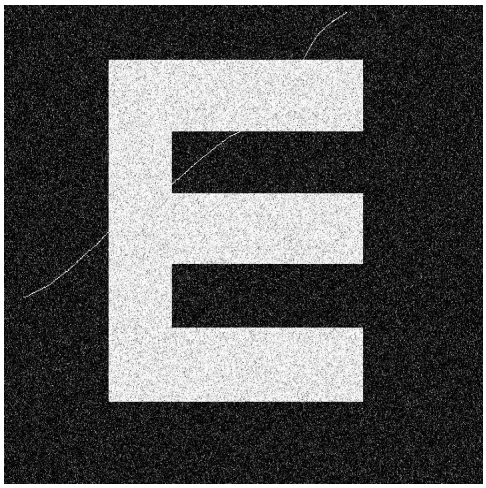

Figure 5. The noised image of the letter 'E' with some added scratches.

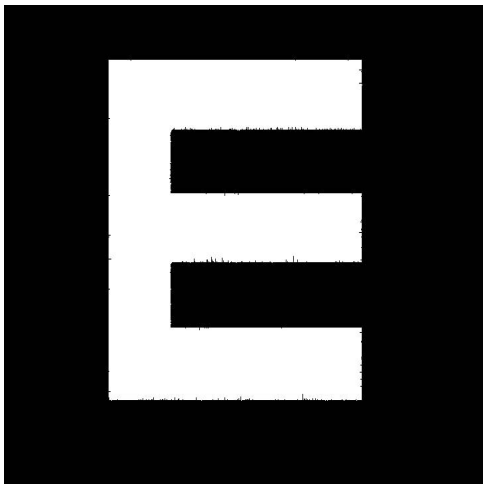

Figure 6. The restored image of the noised image of the letter 'E' with some scratches by applying *Procedure CP*.

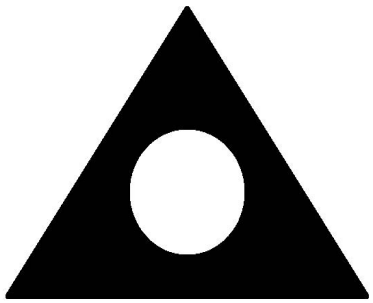

Figure 7. The original image of the second example.

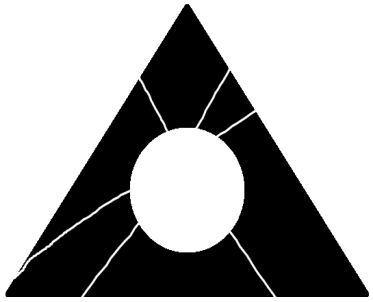

Figure 8. The above image with some scratches.

this image (see Figure 8), and further contaminated it by adding some noise (see Figure 9). As above, we first convert the image displayed in Figure 9 to the image matrix, which has dimensions 544×700. We then apply *Procedure CP* to this image matrix for image retrieval. The restored image is displayed in Figure 10, which shows satisfactory performance of *Procedure CP*.

V. CONCLUSION

In this paper, we tackle the image retrieval problem from a different angle. By converting the problem into a multiple change-point detection problem, with the help of *Algorithm WSW*, we propose *Procedure CP* for restoring a noise contaminated black white image with scratches. As demonstrated in two examples, the new method has satisfactory performance in terms of the quality of retrieved images.

We remark that the algorithm given in [16] can be replaced by other multiple change-point detection methods. Even though, we only consider image retrieval for black white images, it may be extended to restore a corrupted gray

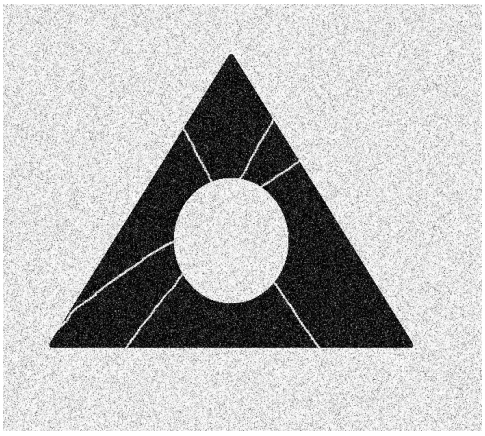

Figure 9. The noised image of the image displayed in Figure 8.

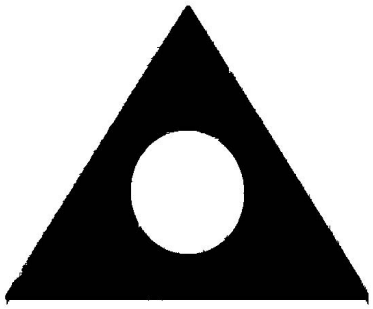

Figure 10. The restored image of the image displayed in Figure 9 by applying *Procedure CP*.

image. The idea used in this paper, i.e., converting the image retrieval problem to a multiple change-point detection problem, may also be applied in other areas such as fast and secure information transmission.

ACKNOWLEDGMENT

This research is partially supported by the Canadian National Science and Engineering Research Council.

REFERENCES

[1] J. Chen and A. K. Gupta, "Parametric Statistical Change Point Analysis. With Applications to Genetics, Medicine, and Finance," 2nd Edition, Boston: Birkhäuser, 2012.

[2] M. Csörgő and L. Horváth, "Limit Theorems in Change-Point Analysis," Chichester: Wiley, 1997.

[3] M. Haidekker, "Advanced Biomedical Image Analysis," Hoboken: Wiley, 2010.

[4] Z. Hlávka, M. Hušková, C. Kirch, and S. G. Meintanis, "Monitoring changes in the error distribution of autoregressive models based on Fourier methods," Test, vol. 21, 2012, pp. 605-634.

[5] M. Hušková and S. G. Meintanis, "Change point analysis based on empirical characteristic function," Metrika, vol. 63, 2006, pp. 145-168.

[6] M. Hušková and S. G. Meintanis, "Change point analysis based on empirical characteristic function of ranks," Sequential Analysis, vol. 25, 2006, pp. 421-436.

[7] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," second ed., Prentice Hall, 2002.

[8] B. K. Gunturk, "Fundamentals of image restoration" in Image Restoration: Fundamentals and Advances, B. K. Gunturk and X. Li, Eds. Boca Raton: CRC Press, 2012, pp. 25-62.

[9] C. Inclán and G. Tiao, G, "Use of cumulative sums of squares for retrospective detection of change of variance," J. Amer. Statist. Assoc., vol. 89, 1994, pp. 913-923.

[10] B. Jähne, "Digital Image Processing," Berlin: Springer, 2005.

[11] X. Li, "Image denoising: Past, present, and future," in Image Restoration: Fundamentals and Advances, B. K. Gunturk and X. Li, Eds. Boca Raton: CRC Press, 2012, pp. 1-24.

[12] E. S. Page, "A test for a change in a parameter occurring at an unknown point," Biometrika, vol. 42, pp. 523527.

[13] G. Qian, X. Shi, and Y. Wu, "A statistical test of change-point in mean that almost surely has zero error probabilities," Aust. N. Z. J. Stat., vol. 55, 2014, pp. 435-454.

[14] P. van Beek, Y. Su, and J. Yang, "Image denoising and restoration based on nonlocal means" in Image Restoration: Fundamentals and Advances, B. K. Gunturk and X. Li, Eds. Boca Raton: CRC Press, 2012, pp. 89-114.

[15] G. Wang and S. Wang, "Recursive computation of Tchebichef moment and its inverse transform," Pattern Recognition, vol.39, 2006, pp. 47-56.

[16] D. Wei, X. Shi, and Y. Wu, "An algorithm for detecting multiple change-points in distribution," preprint.

 ISBN: 978-1-61208-358-2