

Multi-verse Optimization Clustering Algorithm for Binarization of Handwritten Documents



Mohamed Abd Elfattah, Aboul Ella Hassanien, Sherihan Abuelenin and Siddhartha Bhattacharyya

Abstract Binarization process of images of historical manuscripts is considered a challenge due to the different types of noise that are related to the degraded manuscripts. This paper presents an automatic clustering algorithm for binarization of handwritten documents (HD) based on multi-verse optimization. The multi-verse algorithm is used to find cluster centers in HD where the number of clusters is pre-defined. The proposed approach is tested on the benchmarking dataset used in the Handwritten Document Image Binarization Contest (H-DIBCO 2014). The proposed approach is assessed through several performance measures. The experimental results achieved competitive outcomes compared to the well-known binarization methods such as Otsu and Sauvola.

Keywords Binarization · Handwritten documents · Multi-verse optimizer (MVO) H-DIBCO 2014

1 Introduction

Binarization in document analysis field is considered as an open challenge, since the historical manuscript images suffer from different kinds of noise and any proposed

M. A. Elfattah · S. Abuelenin
Faculty of Computers and Information, Computer Science Department,
Mansoura University, Dakahlia Governorate, Egypt

A. E. Hassanien
Faculty of Computers and Information, Cairo University, Giza, Egypt

S. Bhattacharyya (✉)
Department of Computer Application, RCC Institute of Information Technology,
Kolkata 700015, India
e-mail: dr.siddhartha.bhattacharyya@gmail.com
URL: <http://www.egyptscience.net>

M. A. Elfattah · A. E. Hassanien · S. Bhattacharyya
Scientific Research Group in Egypt (SRGE), Cairo, Egypt

systems, such as optical character recognition (OCR) and word spotting (WS), need the proper binarized image, while the accuracy of these systems affected directly with this process (binarization). Binarization is the process of extracting the text without any noise (black) and background (white) [1]. With more kinds of noise as; smudge, multi-colored, bleed-through, unclear background, shadow, broken character. The current systems depend on the binarization as the first step which is affected by noise. These systems are included in many applications such as handwritten recognition, watermarking, and data hiding [2].

Thresholding approaches can be either local or global. In the case of degraded images, global approaches do not perform well [3]. Otsu [4], Kapur et al. [5], and Kittler and Illingworth [6] are considered global methods, while Niblack [7], Sauvola and Pietikäinen [8], and Bernsen [9] are considered local methods. From the literature, many different approaches are presented to binarize the degraded images. However, the binarization process is still an open challenge [10].

Recently, meta-heuristic optimization algorithms have a wide range of applications such as feature selection, image processing, and others. Nature-inspired algorithms are well-known optimization algorithms. In these algorithms, the local optima problem can be solved by sharing the information between candidates [11]. Therefore, in this paper, a new cluster algorithm is proposed using one of the recent optimization algorithms named multi-verse optimizer (MVO) [11]. This algorithm is proposed to address the binarization process of historical documents.

The rest of this paper is organized as follows: Section 2 introduces the basics of MVO algorithm. Section 3 presents the proposed approach. In Sect. 4, the experimental result and discussion are clarified. Finally, conclusions and future works are presented in Sect. 5.

2 Preliminaries: Multi-verse Optimizer (MVO)

MVO is a recent nature-inspired algorithm proposed by Mirjalili et al. [11]. It is based on the three concepts of cosmology (white hole, black hole, and wormhole). The exploration phase is based on (white, black hole), while the wormhole is employed for improving the quality in the exploitation phase [11].

At each iteration, these universes are sorted depending on their inflation rate. The roulette wheel is employed for the selecting to have a white hole:

$$\mathbf{U} = \begin{bmatrix} y_1^1 & \dots & y_1^v \\ \dots & \dots & \dots \\ y_n^1 & \dots & y_n^v \end{bmatrix} \quad (1)$$

where the number of parameters (variables) is presented by v and the number of universes by n .

$$y_i^j = \begin{cases} y_k^j & r1 < NI(Ui) \\ y_i^j & r1 \geq NI(Ui) \end{cases} \quad (2)$$

where y_i^j denotes j th of i th universes. Ui presents the i th universe. The normalized inflation rate is presented by $NI(Ui)$ of the i th universe, $r1$ is a random value in $[0, 1]$, and y_k^j presents the j th parameter of k th universe chosen by a roulette wheel selection mechanism [11].

To update the solutions, the two parameters Traveling Distance Rate (TDR) and Wormhole Existence Probability (WEP) are calculated based on Eqs. 3 and 4:

$$WEP = min + l \times \left(\frac{max - min}{L} \right) \quad (3)$$

The minimum and maximum are presented by min (0.2) and max (1) as in Table 1, respectively, while the current iteration presented by l and L denotes the maximum number of iterations:

$$TDR = 1 - \left(\frac{l^{1/p}}{L^{1/p}} \right) \quad (4)$$

The exploitation accuracy is presented by p . The large value of p indicates high perfect of local search/exploitation. The position of solutions is updated based on Eq. 5:

$$y_i^j = \begin{cases} \begin{cases} Y_j + TDR \times ((ub_j - lb_j) \times r4 + lb_j) & r3 < 0.5 \\ Y_j - TDR \times ((ub_j - lb_j) \times r4 + lb_j) & r3 \geq 0.5 \end{cases} & r2 < WEP \\ y_i^j & r2 \geq WEP \end{cases} \quad (5)$$

where Y_j denotes the j th parameter of the best universe; lb_j and ub_j denote the lower and upper bound of j th variable, while, $r2$, $r3$, and $r4$ are random numbers in $[0, 1]$. y_i^j denotes the j th parameter of i th universe. TDR and WEP are coefficients [11].

3 The Proposed Binarization Approach

Starting with applying the MVO algorithm on the degraded manuscripts image to find the optimal cluster center based on objective function given in Eq. 6 as in the basic K -means clustering algorithm [12]. Depending on the obtained cluster centers

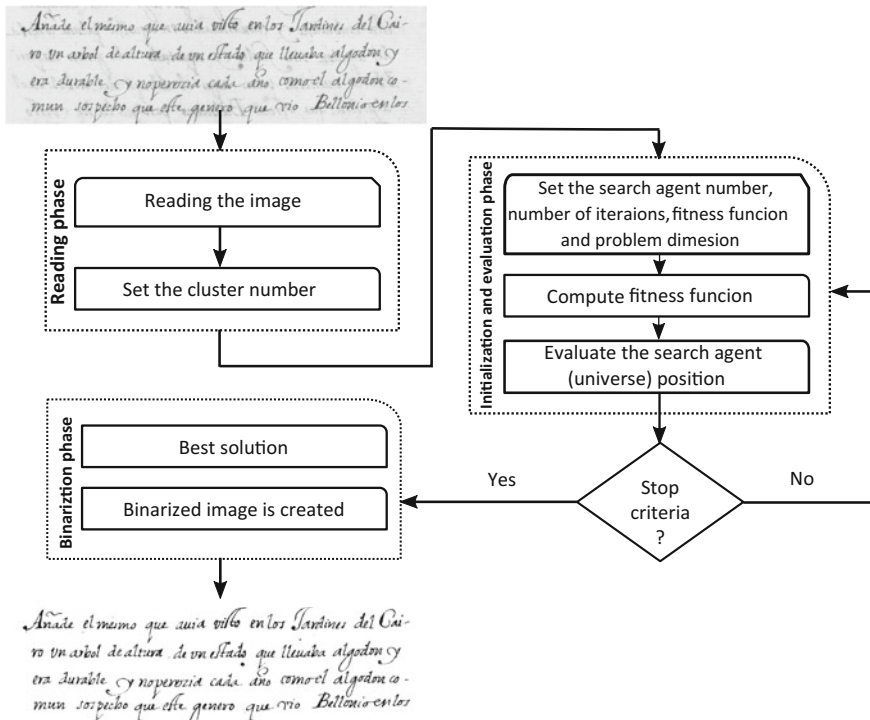


Fig. 1 General architecture of the proposed binarization approach

to create BW (white, black) representing the foreground by white pixels where the darkest cluster denotes the text. In fact, at every iteration, each (universe) search agent updates its position according to (the best position). Finally, the cluster centers are updated, and the binary image is created. Figure 1 illustrates the general architecture of the proposed binarization approach and its phases.

3.1 Fitness Function and MVO Parameters

Equation 6 is the squared error function that is used as an objective function of the multi-verse optimization algorithm typically as in the *k*-means clustering [12]:

Table 1 MVO parameter's setting

Parameter	Value
Number of universes	8
Max iterations	10
Min (Eq. 3)	0.2
Max (Eq. 3)	1
No. of run	30
No. of clusters	2
Dimension	1
Range	[0 255]

$$J = \sum_{j=1}^k \sum_{i=1}^x \|x_i^{(j)} - c_j\|^2 \quad (6)$$

The distance measure among the cluster center c_j and data points $x_i^{(j)}$ is presented by $\|x_i^{(j)} - c_j\|^2$. It denotes the distance of the n data points from their cluster centers.

The foremost target of MVO is to minimize this function. Each cluster is presented within a single centroid. Each universe presents one solution, and its position is updated according to (best solution). For any optimization algorithm, we primarily require setting some parameters value that provides better performance of the proposed approach. Table 1 refers to the MVO parameters setting.

4 Experimental Results and Discussion

H-DIBCO 2014 dataset [13] is used and employed to evaluate the proposed approach. This dataset contains ten handwritten images with different kinds of noise which are collected from *tranScriptorium* project [14]. This dataset is available with its ground truth. This dataset contains illustrative degradations such as bleed-through, faint characters, smudge, and low contrast.

To evaluate the proposed approach, different performance measures [15] are used and employed including F-measure [16, 17], Negative Rate Metric (NRM) [18], Peak Signal-to-Noise Ratio (PSNR) [17], Distance Reciprocal Distortion (DRD) [2], and Misclassification Penalty Metric (MPM) [18, 19]. The high values of F-measure, PSNR, and low value on DRD, NRM, and MPM indicate the best result. In addition, visual inspection is used.

Table 2 Result of MVO on H-DIBCO 2014

Image name	F-measure	PSNR	DRD	NRM	MPM
H01	91.79	20.47	2.05	0.05	0.08
H02	89.48	17.56	2.93	0.06	0.45
H03	98.09	24.04	0.92	0.01	0.03
H04	94.94	18.32	1.60	0.04	0.37
H05	94.38	17.47	2.04	0.04	0.66
H06	94.05	17.56	2.36	0.04	0.51
H07	84.85	15.32	5.98	0.04	7.97
H08	94.12	24.59	1.51	0.04	0.14
H09	88.77	16.99	2.72	0.09	0.21
H10	89.66	17.30	2.51	0.08	0.28
Average	92.01	18.96	2.46	0.04	1.07

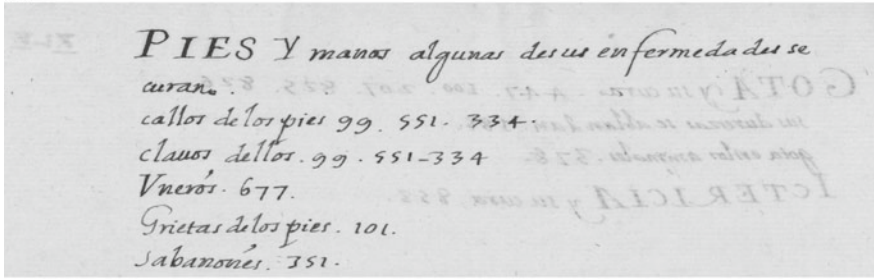
Table 3 Results of the MVO with the state-of-the-art methods on H-DIBCO 2014

Approach name	F-measure	PSNR	DRD
1 [13]	96.88	22.66	0.902
2 [13]	96.63	22.40	1.001
3 [13]	93.35	19.45	2.194
4 [13]	89.24	18.49	4.502
Otsu [13]	91.78	18.72	2.647
Sauvola [13]	86.83	17.63	4.896
MVO result	92.01	18.96	2.46

Table 2 presents the results of MVO on H-DIBCO 2014; the high PSNR value appears in H08 with value 24.59, while the worst value is in H07 with value 15.32. The higher value of F-measure (98.09) is in H03, while the worst is in H07 (84.85). In addition, the better DRD value is in H03 (0.92). The best NRM and MPM appear in H03 (0.01, 0.03), respectively.

Table 3 summarizes the comparison between the approaches submitted to H-DIBCO 2014 competition [13] and the proposed MVO algorithm. According to Table 3, the numbers (1 to 4) indicate the rank of submitted methods with their values of F-measure, PSNR, and DRD. The result of MVO is better than the well-known methods (Otsu and Sauvola) and the method number (4) in all performance measures.

(a)



(b)

PIES Y manos algunas de sus enfermedades se curan.
callos de los pies 99. 551. 334.
clavos de los. 99. 551-334
Vneros. 677.
Grietas de los pies. 101.
Sabanones. 751.

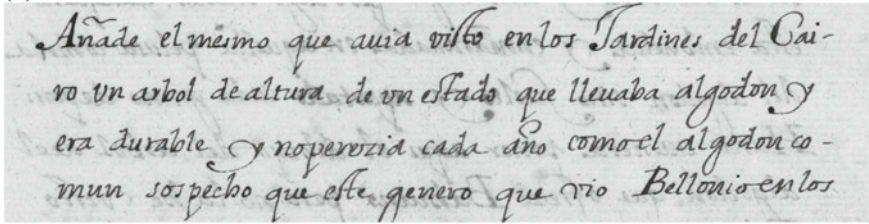
(c)

PIES Y manos algunas de sus enfermedades se curan.
callos de los pies 99. 551. 334.
clavos de los. 99. 551-334
Vneros. 677.
Grietas de los pies. 101.
Sabanones. 751.

Fig. 2 a H08 (H-DIBCO 2014) test sample, b GT, and c MVO result

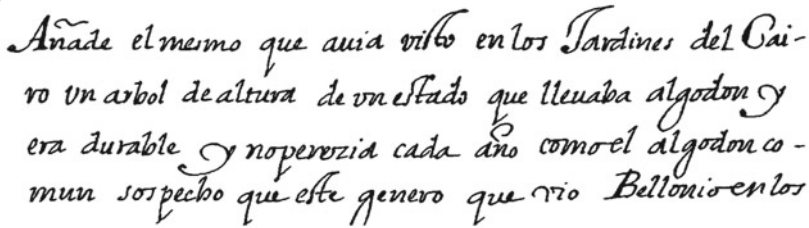
For visual inspection, two images are selected named H08 and H10 as shown in Figs. 2 and 3. Figures 2 and 3 show the comparison between the ground truth images, as shown in Figs. 2b and 3b, and the MVO output images (Figs. 2c and 3c). From these figures, the output images are very close to the ground truth images with complete character structure, but we found some simple noise.

(a)



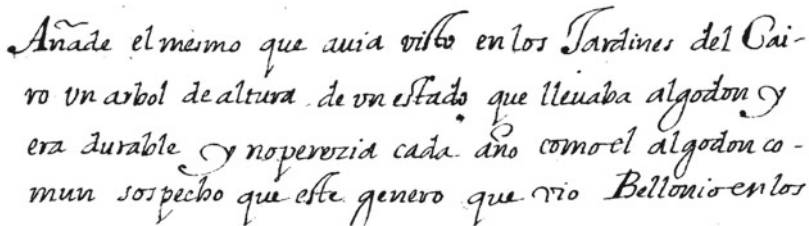
*Añade el mismo que auia visto en los Jardines del Cai-
ro un arbol de altura de vn estado que lleuaba algodón y
era durable y no perozia cada año como el algodón co-
mun sospecho que este genero que vio Bellonio en los*

(b)



*Añade el mismo que auia visto en los Jardines del Cai-
ro un arbol de altura de vn estado que lleuaba algodón y
era durable y no perozia cada año como el algodón co-
mun sospecho que este genero que vio Bellonio en los*

(c)



*Añade el mismo que auia visto en los Jardines del Cai-
ro un arbol de altura de vn estado que lleuaba algodón y
era durable y no perozia cada año como el algodón co-
mun sospecho que este genero que vio Bellonio en los*

Fig. 3 a H10 (H-DIBCO 2014) test sample, b GT, and c MVO result

The convergence rate is the last judgment measure to evaluate the proposed binarization approach. In each iteration, the solution with the best fitness is kept and it is used to create the convergence curves as in Fig. 4. This figure presents the convergence curve for two different images, and the lower fitness value with increasing the number of iterations demonstrates the convergence of the proposed approach. It is also remarkable that the fitness value decreased dramatically. The optimization problem here is a minimization problem. We can conclude from this figure that the MVO is a promising approach to address the binarization problem.

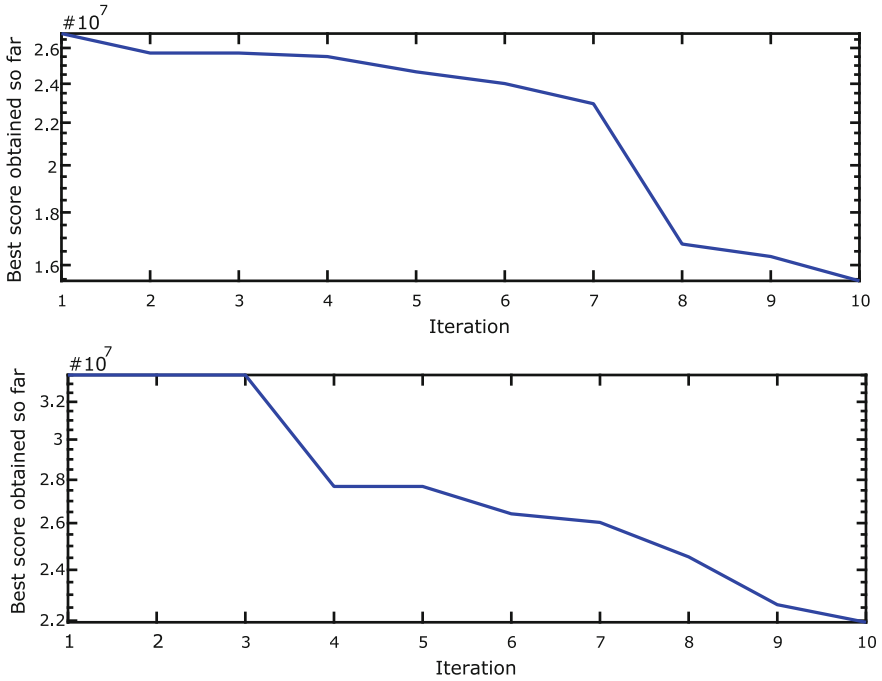


Fig. 4 MVO convergence curve

5 Conclusions and Future Works

This paper presents a binarization approach based on the MVO algorithm, which is employed for minimizing the distance between clusters. The convergence curve rate proves the high speed of MVO algorithm. This approach can deal with various kinds of noise.

As future work, it is planned to use preprocessing phase which can improve the accuracy of binarization. Furthermore, hybridization with other optimization algorithms will be used to improve the results in [20–24]. A comparative analysis between the basic MVO and a chaotic version of it based on different chaos maps and different objective functions will be presented to improve the OCR recognition rate in [25, 26] by using it in the binarization phase.

References

1. Mesquita RG, Silva RM, Mello CA, Miranda PB (2015) Parameter tuning for document image binarization using a racing algorithm. *Expert Syst Appl* 42(5):2593–2603
2. Lu H, Kot AC, Shi YQ (2004) Distance-reciprocal distortion measure for binary document images. *IEEE Signal Process Lett* 11(2):228–231
3. Singh BM, Sharma R, Ghosh D, Mittal A (2014) Adaptive binarization of severely degraded and non-uniformly illuminated documents. *Int J Doc Anal Recognit (IJ DAR)* 17(4):393–412
4. Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 9(1):62–66
5. Kapur JN, Sahoo PK, Wong AK (1985) A new method for gray-level picture thresholding using the entropy of the histogram. *Comput Vis Graph Image Process* 29(3):273–285
6. Kittler J, Illingworth J (1986) Minimum error thresholding. *Pattern Recognit* 19(1):41–47
7. Niblack W (1985) An introduction to digital image processing. Strandberg Publishing Company
8. Sauvola J, Pietikäinen M (2000) Adaptive document image binarization. *Pattern Recognit* 33(2):225–236
9. Bernsen J (1986) Dynamic thresholding of grey-level images. *Int Conf Pattern Recognit* 2:1251–1255
10. Hadjadj Z, Cheriet M, Meziane A, Cherfa Y (2017) A new efficient binarization method: application to degraded historical document images. *Signal Image Video Process* 1–8
11. Mirjalili S, Mirjalili S, Hatamlou A (2016) Multi-verse optimizer: a nature-inspired algorithm for global optimization. *Neural Comput Appl* 27(2)
12. MacQueen J et al (1967) Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol 1, pp 281–297
13. Ntirogiannis K, Gatos B, Pratikakis I (2014) ICFHR2014 competition on handwritten document image binarization (h-dibco 2014). In: *2014 14th international conference on frontiers in handwriting recognition (ICFHR)*. IEEE, pp 809–813
14. <http://transcriptorium.eu>
15. Gatos B, Ntirogiannis K, Pratikakis I (2009) ICDAR 2009 document image binarization contest (DIBCO 2009). In: *10th international conference on document analysis and recognition, 2009 (ICDAR'09)*. IEEE, pp 1375–1382
16. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. *Inf Process Manag* 45(4):427–437
17. Ntirogiannis K, Gatos B, Pratikakis I (2013) Performance evaluation methodology for historical document image binarization. *IEEE Trans Image Process* 22(2):595–609
18. Pratikakis I, Gatos B, Ntirogiannis K (2010) H-dibco 2010-handwritten document image binarization competition. In: *2010 international conference on frontiers in handwriting recognition (ICFHR)*. IEEE, pp 727–732
19. Young DP, Ferryman JM (2005) Pets metrics: on-line performance evaluation service. In: *Joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance (VS-PETS)*, pp 317–324
20. Elfattah MA, Abuelenin S, Hassanien AE, Pan JS (2016) Handwritten arabic manuscript image binarization using sine cosine optimization algorithm. In: *International conference on genetic and evolutionary computing*. Springer, pp 273–280
21. Mostafa A, Fouad A, Elfattah MA, Hassanien AE, Hefny H, Zhu SY, Schaefer G (2015) Ct liver segmentation using artificial bee colony optimisation. *Procedia Comput Sci* 60:1622–1630
22. Mostafa A, Elfattah MA, Fouad A, Hassanien AE, Hefny H (2016) Wolf local thresholding approach for liver image segmentation in ct images. In: *Proceedings of the second international Afro-European conference for industrial advancement (AECIA 2015)*. Springer, pp 641–651
23. Ali AF, Mostafa A, Sayed GI, Elfattah MA, Hassanien AE (2016) Nature inspired optimization algorithms for ct liver segmentation. In: *Medical imaging in clinical applications*. Springer, pp 431–460

24. Hassanien AE, Elfattah MA, Aboulenin S, Schaefer G, Zhu SY, Korovin I (2016) Historic handwritten manuscript binarisation using whale optimisation. In: 2016 IEEE international conference on systems, man, and cybernetics (SMC). IEEE, pp 003842–003846
25. Sahlol AT, Suen CY, Zawbaa HM, Hassanien AE, Elfattah MA (2016) Bio-inspired bat optimization algorithm for handwritten arabic characters recognition. In: 2016 IEEE congress on evolutionary computation (CEC). IEEE, pp 1749–1756
26. Sahlol A, Elfattah MA, Suen CY, Hassanien AE (2016) Particle swarm optimization with random forests for handwritten arabic recognition system. In: International conference on advanced intelligent systems and informatics. Springer, pp 437–446