J Inno Sci Eng,2022, 6(2):233-247
https://doi.org/10.38088/jise.1135756

Research Article

An Application of Non-Dominated Sorting Genetic Algorithm for Reversible Data Hiding Based on Histogram Shifting in Neuroimages

Fusun Er [1*], Yildiray Yalman [2]

[1] *Pîrî Reis University,Department of Information Systems Engineering, 34940, Turkey*
[2] *Pîrî Reis University,Department of Information Systems Engineering, 34940, Turkey*

Cite this paper as: Fusun, E. and Yildiray, Y. (2022). An application of non-dominated sorting genetic algorithm for reversible data hiding based on histogram shifting in neuroimages. 6(2):233-247

*Corresponding author: Fusun Er
E-mail: fer@pirireis.edu.tr

Received Date: 25/06/2022
Accepted Date: 26/08/2022

Availableonline at
http://jise.btu.edu.tr/

Abstract

This paper presents an application of a multi-objective non-dominated sorting genetic algorithm with a modified chromosome encoding for histogram shifting-based multiple reversible data hiding scheme in neuroimages which aims to minimize distortion and maximize capacity. The modified chromosomes encoding scheme is designed according to the zero-bin characteristic of the intensity histogram of the structural magnetic resonance imaging scans of the human brain. A detailed experimental study has been carried out for assessing the effect of non-dominated sorting for multi-objective optimization compared to Euclidian distance, the convenience of modified chromosome encoding scheme for medical images compared to non-medical images. The performance of the proposed method has been measured in terms of the peak signal-to-noise ratio (PSNR) for image quality and the bits per pixel (bpp) for capacity assessments. The experimental results show that the proposed method is better than its counterparts.

Keywords: Genetic algorithm, information security, magnetic resonance imaging, reversible data hiding

1. Introduction

Data hiding is the process of sending confidential data covertly embedded into a digital media, such as digital images, video, and signals [1]. The idea behind the utilization of digital images as a carrier in data hiding technologies is based on the inability of people to notice small changes in digital images by naked eye. Many data hiding schemas have been proposed in the literature that successfully employed various kinds of images in several application areas, including medical images. The Health Insurance Portability and Accountability Act (HIPAA) of 2003 and 2005 include a set of privacy and security rules that forces medical professionals and institutions to ensure patient confidentiality and privacy, even in digital media [2]. Since medical images contain critical information for diagnosis, the applied data hiding technique must ensure the reversibility of the cover image. Nowadays, it is of great interest to study data hiding techniques to hide identification information inside magnetic resonance images for establishing more secure data transmission channels. Most of the existing reversible data hiding schemes are based on histogram shifting (HS) , which was initially proposed by Ni et al. [1]. Kurnaz et al. proposed an histogram-shifting based method that does not require shifting while preserving the visual quality of stego images [3]. A typical HS-based data hiding scheme start by selecting one or more pairs of peak and zero bins in the histogram of a cover image. Then, it shifts the bins between peak bin and zero bin by one toward the zero bin. The main purpose is to create a gap for hiding a secret message in the size of the frequency of the peak bin, which the frequency of the selected bin determines the hiding capacity. On the other hand, the image distortion depends on the total number of shifted pixels. In a single-pair HS-based scheme, the peak bin with the highest frequency is matched to its nearest zero bin, thus, the number of pixels affected by shifting is the smallest [4]. A multiple embedding scheme should be considered in order to achieve a higher capacity, which consecutively employs more than one embedding procedure on pairs of different peaks and zero bins [5]. If it is aimed to increase the embedding capacity despite the high distortion in embedded images, the same pair selection procedure for a single pair can be repeated for multiple HS-based schemes by matching the next highest peak bin with its closest zero bin in case of a small number of zero bins existing in a histogram. However, multiple HS-based data hiding is a non-deterministic polynomial-time (NP)-hard problem when the size of the solution space is too large to find the best mapping of peak and zero bins. This problem is known as the rate-distortion optimization problem. Several algorithms have been proposed in the literature to find an optimal set of peak and zero pairs [6], [7] for multiple pair histogram shifting. Tian's algorithm searches for redundancy in digital images to achieve high embedding capacity while keeping the distortion low [8]. In practice, when the solution space is too large to find an optimal solution in a reasonable time, heuristic algorithms [9] are recommended to search for an optimal solution [10]. Recently, Wang et al. proposed a genetic algorithm-based embedding scheme in order to automatically determine the number of peak and zero bin pairs and their corresponding values [11], [12] and [13]. Furthermore, dynamic programming-based reversible data hiding algorithms were proposed to solve the optimization problem [14], [15].

This study aimed to implement a reversible and high-capacity HS-based data hiding technique for magnetic resonance imaging (MRI), which is one of the most widely used medical imaging tools. MRI is a non-invasive technology that produces detailed anatomical and functional images of the inner body without any exposure to radiation. Thus, a genetic algorithm model is proposed in this paper for solving a rate-distortion optimization problem for HS-based multiple reversible data embedding schemes.

The paper is organized as follows: Section 2 provides background information on the medical imaging techniques that were used to obtain the data that are included in this study. Section 3 describes the dataset used to evaluate the proposed model. The framework for rate-distortion optimization on HS-based data hiding is demonstrated in sub-section 3.2. Then, experimental test results are presented and discussed. Final remarks are summarized in the Conclusion Section..

2. Background and Motivation

This section provides brief information about in vivo imaging of the human brain and how its intensity histogram characteristics inspired the chromosome structure of the proposed genetic algorithm (GA) model.

A. Structural imaging of the human brain

The human brain is the most complex organ of the human body and a part of the central nervous system (CNS). The brain is composed of two types of tissue: white matter (WM) and grey matter (GM). In addition, cerebrospinal fluid (CSF) is a clear plasma-like fluid that fills the brain ventricles. Grey matter contains relatively fewer myelinated neurons compared to white matter, which is mainly distributed on the surface of the brain cortex. The white matter appears white due to the abundance of the fatty substance (myelin) in its structure [16]. Magnetic resonance imaging is a widely used in vivo imaging technique for studying the human brain. Different tissue types have different longitudinal (or spin-lattice) relaxation times (T1) that is a measure of the time taken for spinning protons to recover about 63% of the magnetization along the longitudinal direction. T1-weighted (T1-w) imaging is one of the basic pulse sequences in MRI, which exhibits contrast differences between different tissue types. The excited hydrogen nuclei in fat recover more rapidly along the longitudinal axis; thus, regain much longitudinal magnetization during the repetition time (TR) interval. Due to the lipid composition of myelin, white matter has higher intensity values and appears lighter than grey matter in T1-weighted images of the healthy human brain. The lowest signal intensity is obtained from protons in the water molecules of CSF. Hereby, CSF appears black in T1-w imaging of the brain [17].

B. Motivation

The histogram of T1-weighted imaging of the healthy human brain is characterized by three main mounds that correspond to three main tissue types: cerebrospinal fluid, grey matter, and white matter, respectively [17]. In this study, a group of non-zero neighboring bins delimited by zero bins of the histogram is called a “cluster", and similarly, a group of zero neighboring bins delimited by non-zero bins is called a “gap". A typical intensity histogram of a structural MRI image has either narrow gaps between thick clusters that create many candidate peak points per zero points or wider gaps between thin clusters that form many unmatched zero locations. This characteristic is the main motivation behind the algorithmic design of the presented algorithm. The chromosome structure of this study is designed to encode the selection of zero-bin pairs for histogram shifting, such that adjacent zero-bins within a gap are matched with the non-zero bins of the neighboring cluster.

C. Materials and Methods

The efficiency of the proposed Genetic Algorithm (GA) is presented on axial slices of a set of T1-weighted imaging scans belonging to healthy elderly. Additionally, the proposed approach is compared with a non-heuristic schema. Furthermore, the proposed schema is tested for widely used non-medical images: Lena, Baboon, and Peppers.

Experimental Dataset

A total of 30 healthy elderly subjects have been included in this study. For each subject, their spatially normalized and skull-stripped anatomical MRI brain scans were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data archive. The scans are three-dimensional 16-bit depth magnetic resonance images with a size of 110×110×110 and a resolution of 2 mm×2 mm×2 mm. The experiments were performed on the 60th axial slice of the volumes at a size of 110×110. Fig.1 presents an axial slice of one of the MRI scans used in the experimental evaluation of the proposed model.

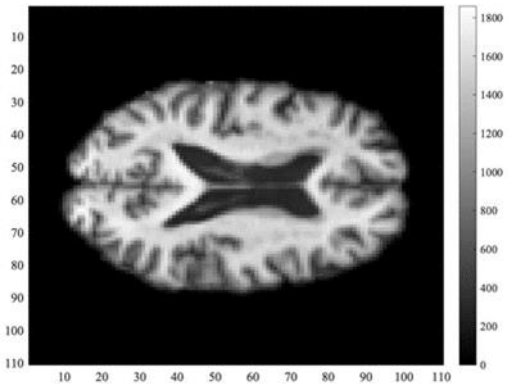


Figure 1. An axial slice of the dataset used in the experiments.

Proposed model for reversible data-hiding

The conventional phases of a genetic algorithm have been performed, which includes (1) initial population generation, (2) parent selection, (3) crossover, (4) mutation, and (5) next population selection. The overall GA schema requires three system parameters and one input: the size of the population (n_p), the number of epochs (n_e) and the mutation probability (r_m) are the system parameters; the histogram model of the cover image (H) is the input parameter. The aim of this GA model is to find an optimum set of zero-nonzero bin pairs based on the histogram model of a particular cover image. In this study, the model proposes a new chromosome encoding scheme based on the histogram structure, and employs a non-dominated sorting algorithm to find the best individual for the next population.

Chromosome encoding based on the intensity histogram of the image

The proposed chromosome structure is highly related to the histogram model of the cover image. The data structure of a histogram model (H) is composed of gaps (G) and clusters (C) denoted as $H = (G, C)$, where the number of gaps is represented by $|G| = n_g$ and the number of clusters is represented by $|C| = n_c$.

G is a sorted sequence of gaps that are labelled as g_i in such a way that $G =< g_1, \ldots., g_{n_g} >$. Each gap g_i

corresponds to one of the zero-bin gaps of the histogram in ascending order that g_m represents bins with smaller numbered bins than what g_n represents when $m < n \leq n_g$. Thus, a single gap g_i is implemented as a sequence of consecutive zero bin-numbers (ZB_i). The number of zero-bins in a gap is at least one, but the size may vary according to the histogram of the image ($|ZB_i| > 0$).

Similarly, C is used to represent all clusters of sequential non-zero bin-numbers $C =< c_1, \ldots., c_{n_c} >$ in ascending order. A single cluster c_i has one or more non-zero bins associated with its intensity as a sequence of pairs (NZB_i) in opposite to gaps since all intensities are zero in gaps. Thus, NZB_i represents non-zero bins of the cluster c_i as $NZB_i =\ll b_{i1}, q_{i1} >, \ldots., < b_{ik}, q_{ik} >>$, where b's are the bin numbers and q's are the corresponding intensities of the bin numbers such that $k = |NZB_i|>0$ is the number of non-zero bins in the clusterc_i.

Since, we included the bin number 0 (zero) that does not exist in the ROI of the image and the bin number "one plus the highest intensity value of the image", a histogram model starts and ends with a gap, which yields $|G| = n_g$ that is always equals to ($|C| = n_c$) + 1 and the bin numbers of g_i is always followed by the bin number of c_i. The following figure Fig.2 illustrates a symbolic histogram of an arbitrary image with a size of 10×10 and with intensity levels in the range of [1−36]. The bin numbers zero and 37 (one more than the highest intensity) are artificially added to the histogram to guarantee that it will start and end with a gap. Therefore, the sample histogram has four gaps ($n_g = 4$) and three clusters ($n_c = 3$). The gaps have three, six, six and three bin numbers in such that $ZB_1 =< 0,1,2 >$, $ZB_2 =< 9,10,11,12,13,14 >$, $ZB_3 =< 23,24,25,26,27,28 >$ and $ZB_4 =< 35,36,37 >$, respectively. The clusters are implemented as two-dimensional tuples that $NZB_1 =\ll 3,5 >, < 4,6 >, < 5,4 >, < 6,3 >, < 7,6 >, < 8,4 >>$,$NZB_2 =\ll 15,2 >, < 16,5 >, < 17,8 >, < 18,7 >, < 19,5 >, < 20,5 >, < 21,3 >, < 22,4 >>$ and $NZB_3 =\ll 29,4 >, < 30,6 >, < 31,5 >, < 32,7 >, < 33,6 >, < 34,5 >>$. As a crosscheck, the sum of intensities of all clusters ($\sum_{i=1}^{n_c} \sum_{j=1}^{k} q_{ik}$) is equals to the number of pixels ($10 \times 10 = 100$).

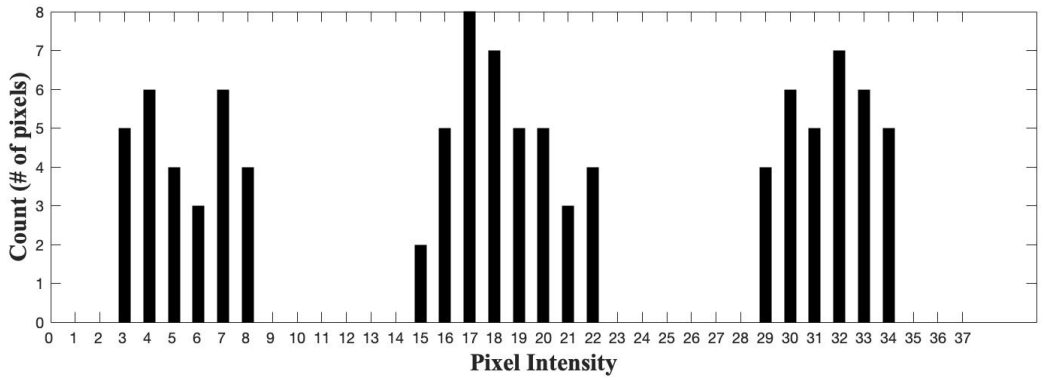


Figure 2. An illustration of an intensity histogram of an arbitrary image with a size of 10x10.

The proposed chromosome encoding scheme based on the histogram model is a series of gene segments that are located in gaps. A gene segment is composed of zero or more encoder-genes and a regulatory gene, as illustrated in Fig.3. Each encoder gene is associated with one of the successive bins of the gap on which it is located. A regulator gene can take a boolean value that expresses (or suppresses) the corresponding encoder-genes. In addition to genes, a gene segment can be either left-handed or right-handed that determines the cluster for matching bin numbers, which is called "encode-direction". If a gene segment of the gap g_i is left-handed then bins of the gap are matched with the bins of the cluster c_i otherwise they match with the cluster c_{i+1}. Each gap may hold zero or one

left-handed gene segment and zero or one right-handed gene segment except the first and the last gaps. The first gap may contain zero or one right-handed gene segment, and the last gap may contain zero or one left-handed gene segment. Therefore, an individual chromosome may have at most $\big(2 \times (n_g - 2)\big) + 1 + 1 = 2 \times n_g - 2 = 2 \times (n_g - 1) = 2 \times n_c$ numbers of gene segments.

There exist some constraints that must be considered during the training of genetic algorithm (GA) model. The first one says that two gene segments can not overlap, therefore the same zero bin will not be paired with more than one nonzero bin. The second one says the total number of encoder-genes of gene-segments that share the same cluster is limited by the number of bins of the cluster. The final one deals with where the gene-segment is placed in the gap, because a shorter distance between the matching points results in a higher PSNR value in the data embedded-image.

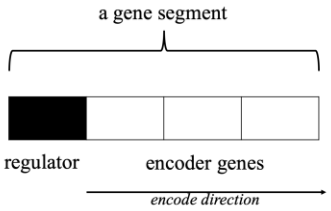


Figure 3. A gene segment

The Fig.4 illustrates an instance of the proposed chromosome structure with three gene segments. The first gene segment is a left-handed gene segment that matches the zero-bins in the range of 9-13 with some non-zero bins of the first cluster. The second gene segment is inactivated. The last gene segment is an active and left-handed segment that has only three zero-bins to match with the bins of the third cluster.

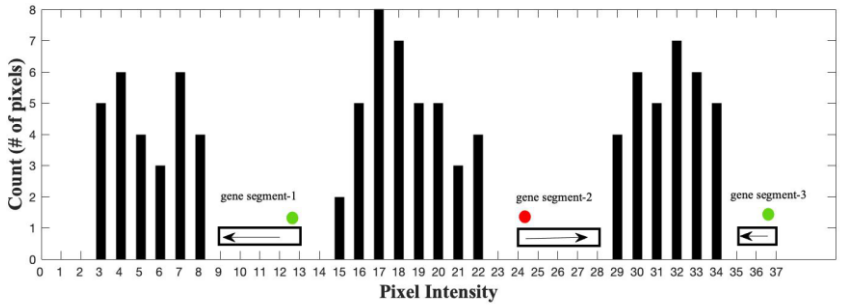


Figure 4. An instance of chromosome encoding on the sample histogram

D. Training of the GA model

The training procedure of the proposed GA model for the given histogram model is given in Algorithm 1. The procedure runs for the given number of epochs (n_e). In each epoch, all individuals in the population are sorted using a non-dominated multi-objective sorting algorithm. The last two-thirds of the population having high scores are randomly paired to generate two offspring. The generated offspring are exposed to mutation with the given mutation probability (r_m). This GA model employs customized implementations of crossover and mutation

according to the proposed chromosome algorithm.

Algorithm 1: Training initial population for n_e numbers of epochs

function GA_TRAIN(n_p , n_e, r_m, H**) :** P
Input: n_p , n_e , r_m are system parameters; H is the histogram model of the image
Output: $P = \{P_i \mid i = 1..n_p\}$ is a sequence of trained individual chromosomes.
/* Initialization */
for $i \leftarrow 1$ *to* n_p
  $P_i \leftarrow randomly\ generated\ chromosome$
end
/* Training */
for $epoch \leftarrow 1$ *to* n_e
  $P \leftarrow \boldsymbol{SORT\_NSA}(P)$
  $Q \leftarrow \emptyset$ // Q will keep generated offspring
  $k \leftarrow \lfloor n_p \times (2/3) \rfloor$ // k is the number parents
  $Ix \leftarrow a\ random\ permutation\ of\ \{n \mid n = 1, 2, .. k\}$
  for $i \leftarrow 1$ *to* k *by* 2
    $offspring \leftarrow \boldsymbol{CROSSOVER}(P_{Ix_i}, P_{Ix_{i+1}})$
    if random_number**(0,1)** $\leq r_m$
      $offspring \leftarrow \boldsymbol{MUTATE}(offspring)$
    end
    $Q \leftarrow Q \cup offspring$
  end
  $P \leftarrow \boldsymbol{SORT\_NSA}\ (P \cup Q)$
  $P = \{P_i \mid i = 1..n_p\}$ // get better n_p number of individuals
end
return P
end function

E. Sub procedures of Training GA Model

E.1. SORT_NSA Procedure: A non-dominated sorting algorithm (NSA) is applied to a population P that consists of n_p number of individuals denoted by $\{P_i \mid i = 1..n_p\}$, assigning each individual to one of the k number of Pareto Fronts $F = \{F_i \mid i = 1..k\}$. The pareto index (i), is the ranking of the pareto front, where the pareto front F_m provides better solutions than F_n for a particular multi-objective optimization problem if $1 \leq m < n \leq k$. On the other hand, the solutions of a Pareto front do not dominate each other. In this study, an iterative approach is followed in the implementation of NSA to sort solutions of the given population aiming to maximize both peak signal-to-noise ratio (PSNR) and data embedding capacity. Initially, all solutions of a population compose a set of remaining solutions. In each iteration of a loop with a counter variable (i) is incremented by one starting from one; the unassigned solutions that are not dominated by any other unassigned solutions are assigned to the Pareto front F_i, until there are no more solutions to assign. Fig.5 illustrates the first three iterations of a NSA procedure for an example population with ten individuals. At the end of the first iteration, two solutions marked with red-cross are selected for the Pareto Front F_1, which are better solutions then all the remaining eight solutions, but they do not dominate each other. One of the red-crossed solutions is better in PSNR and the others offer higher capacities. Then, the second iteration runs for the remaining eight solutions to compose the second Pareto front F_2 in the same way. At the end of the third iteration, only one solution remained, which will be assigned to the last and fourth Pareto front. If sorting within solutions of a Pareto front is needed, then crowd-distance is applied to measure the surrounding density of a solution.

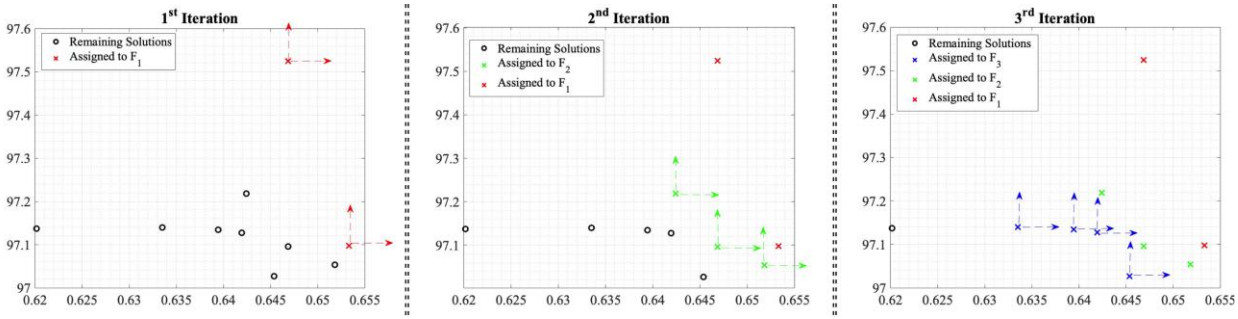


Figure 5. An illustration of non-dominated sorting procedure for a population with ten solutions

The distance value of a particular solution among all solutions in a multi-objective optimization environment is defined as the sum of the differences between the values of the nearest two neighboring solutions when sorted separately according to each objective function. The crowding distance of solutions having the lowest and highest values in an objective function is considered with infinite values so that they are excluded. In this study, solutions with a smaller crowding distance are preferred to achieve optimum solutions in both objectives rather than just one purposive dominant one. This measurement is applied in the selection of the better $n_p$solutions for the next generation as it is in the 18^{th} line of the Algorithm 1 only for the Pareto front of Npth solution.

The Algorithm 2 summarizes the procedure in pseudocode to calculate the crowding distance for all solutions of a particular Pareto front. The algorithm gets individuals (S) in a given Pareto and returns corresponding crowding distance values (CD). Initially all individuals have zero crowd-distance value. Then the algorithm considers each objective separately as an outer loop start in the sixth line. The individuals are sorted according to their j^{th}objective function values. Afterwards, the individuals having the highest and lowest values are excluded with infinite crowded distance values. Finally, the absolute distance of its two neighboring solutions is added to its corresponding crowding distance values in the 11^{th}line.

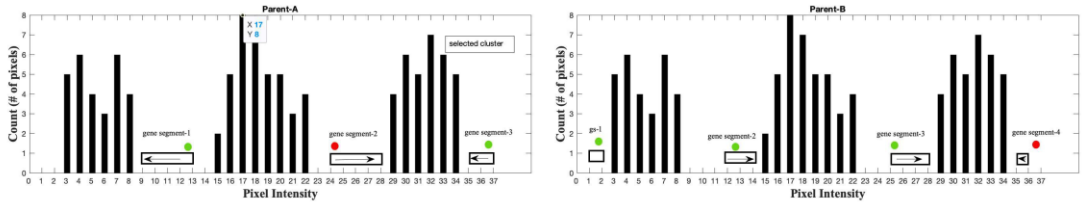
Algorithm 2: Calculate crowding distances of individuals having the Pareto P_r.

function CROWDING_DISTANCE(S) : CD
Input: S= $\{s_i \mid i = 1,2,..,n\}$ is a sequence of n number of solutions in a given Pareto
Output: $CD = \{cd_i \mid i = 1,2,..,n\}$ is a sequence of crowding distances of individuals

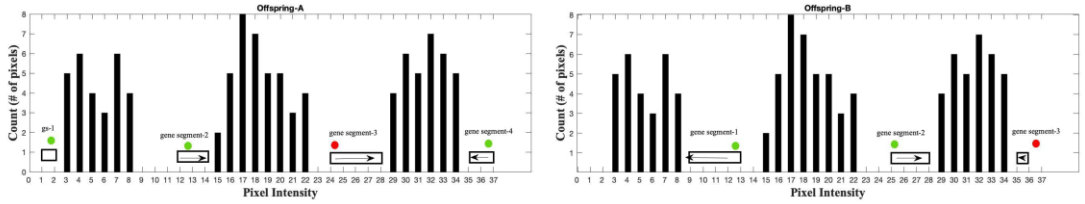
1 /* Initialization */
2 **for** $i \leftarrow 1$ *to* n
3 $cd_i \leftarrow 0$
4 **end**
5 $M = \{m_{i,j} \mid i = 1,2,..,n \text{ and } j = 1..2\}$ // M is a sequence of objective function of n number of solutions in a two-objective involving two objective environment. $m_{i,j}$ stands for the j^{th}objective function value of s_i
6 **for** j $\leftarrow 1$ *to* 2
7 $M',\ T \leftarrow SORT(M_{*,j})$ // T= $\{t_i \mid i = 1,2,..,n\}$ and $M' = \{m'_i \mid i = 1,2,..,n\}$ // in such that $m'_i = m_{t_i,j}$
8 $cd_{t_1} \leftarrow \infty$
9 $cd_{t_n} \leftarrow \infty$
10 **for** k $\leftarrow 2$ *to* $(n-1)$
11 $cd_{t_k} = cd_{t_k} + |m'_{i-1} - m'_{i+1}|$
12 **end**
13 **end**
14 **return** CD
15 **end function**

E.2. CROSSOVER Procedure: As a part of the presented algorithm, the crossover method was designed to preserve the proposed chromosome structure and to comply with the defined constraints. The crossover method starts with a randomly selected twenty-percent of all regulated gene segments of the first parent, which are inherited by the first offspring. The clusters belonging to the remaining clusters of the second parent were transferred to the first offspring (and vice versa for the second offspring). The intersecting encoder genes of the overlapping gene segments are cropped in favor of a randomly chosen one.

Fig.6 illustrates the proposed crossover technique based on the histogram model used in Fig.2. Two different solutions of the histogram (*Parent-A* and *Parent-B*) are crossed over that generated *O spring-A* and *Offspring-B.* The third of the three clusters is selected as crossover point to exchange gene-segments among parents, so that, the *Offspring-A* inherited its first and second gene-segments from *Parent-B* and its third and fourth gene-segments are inherited from *Parent-B.*



a) Parents



b) Offspring

Figure 5. Illustration of a Crossover Operation

E.3. MUTATE Procedure: The mutation affects both regulator genes and encoder-genes. The mutation of regulator genes is a simple procedure in which the regulator genes of randomly selected ten percent of the gene-segments are toggled. Since the previously expressed regulator genes will begin to be suppressed and vice versa, this kind of mutation affects the total capacity. On the other hand, twenty percent of the gene segments were mutated by deleting the encoder-gene next to the regulator gene and entailing it to its complement gene segment. This kind of mutation does not change the capacity. On the other hand, since it affects the matching between zero-bins and peak-bins, such a mutation is related to the distortion.

F. Alternative Distance Measure as opposed to NSA

An Euclidean distance-based fitness value is used to sort individuals in a population as a comparison to NSA in the third and fifteenth lines of Algorithm 1. The fitness value of the chromosome is defined as the magnitude of the vector v in such that $fitness = |\vec{v}| = \langle c, p \rangle^T$, where c and p are scalars in the range of $[0 - 1]$.

The scalar p stands for the ratio of PSNR to the maximum possible PSNR value of the image; and, the scalar c is

the ratio of the capacity in bits to the maximum possible capacity of the image.

In this study, PSNR is calculated for 16-bit data as $PSNR = (20 \times \log_{10}(2^{16} - 1)) - (10 \times \log_{10} MSE)$, where MSE is the mean squared error between the original image and the data embedded-image. The maximum possible PSNR value is calculated by assuming MSE is equal to the epsilon that indicates the minimum distortion occurred by moving only one pixel by one, which is around 135 dB.

3. Results and Discussion

This section presents and discusses experimental results on the dataset explained in Section 2-C.

A. Histograms of the Images of the Experimental Dataset

As it is stated in the background and motivation section, this study is based on the characteristics of the histogram of a structural MRI image. In this section, the intensity histogram of one of the images of the dataset is presented in Figure 7. The histogram has 2125 bins, of which 335 of them are zero-valued bins that are around 16% of the total. Theoretically, the highest 335 peaks can be shifted to all zero locations, but, in practice, only 298 zero-bins are available for shifting due to the characteristics of the histogram.

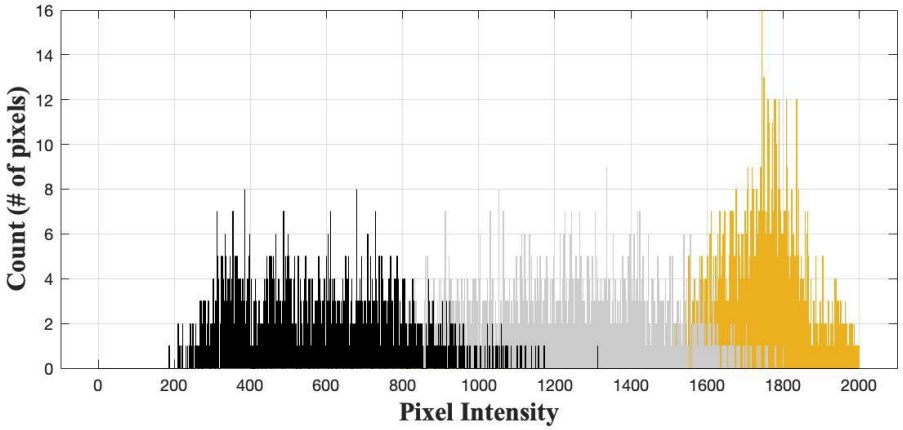


Figure 7. Pixel Intensity Histogram of an Image of the Dataset

The gap-cluster statistics of the histogram are summarized as follows: The wideness statistics of all 231 clusters were 7.74 ±14.09 [1-171] (mean ±std[min-max]). Among 335 zero locations, only 298 of them can be used for shifting, 224 of which can be used for either one of two neighbourhood clusters for shifting. The histogram statistics of the three mounds are given separately in Table 1. The mean signal intensities of cerebrospinal grey matter and white matter are 574, 1228 and 1749, in that order. All three mounds are almost symmetrical bell-shaped ($-0.5 \leq$ skewness ≤ 0.5) but tails are thinner than the normal distribution (kurtosis <3) for cerebrospinal fluid and grey matter.

Table 1: Table explanation.

	Cerebrospinal fluid	Grey matter	White matter
Mean Value	574	1228	1749
Standard Deviation	199	220	122
Median Value	557	1238	1750
Minimum Value	186	744	1411
Maximum Value	1313	1800	2124
Mode	385	1336	1744
25th Percentile	400	1059	1666
75th Percentile	728	1403	1819
Skewness	0.4	-0.2	0.23
Kurtosis	2.52	2.17	3.22

B. Experimental Results

The system parameters are empirically set as n_p = 100 , n_e = 150 and r_m = 10 for the proposed GA model. Fig.8 presents scatter plots of the embedding capacity and PSNR values (in dB) of the individuals of the initial population, the evolved population with the NSA method and the evolved population with the Euclidian distance method; in black, green, and red colors, respectively. The embedding capacity is presented in a rate of embedded data length in bits to the maximum capacity in bits (listed in Table 2). Since, all individuals of the evolved population with the Euclidian distance method provide the solutions with the same capacity and PSNR values, all individuals plotted on top of each other. The mean capacity of the individuals of the initial population is computed to be 0.54 bpp, with a standard deviation of 0.0094. After 150 epochs of training with a procedure using NSA sorting algorithm, the mean PSNR value of the individuals on the first Pareto front of the population was statistically significantly increased compared with the initial population ($p<0.001$), with a mean$\pm$std of 99.82$\pm$0.28, while the capacity value remained same statistically. On the other hand, after 150 epochs of training with a procedure using Euclidian distance sorting algorithm, the capacity value improved significantly ($p<0.001$), with a mean$\pm$std of 0.60$\pm$0.00, but, the PSNR value has decreased.

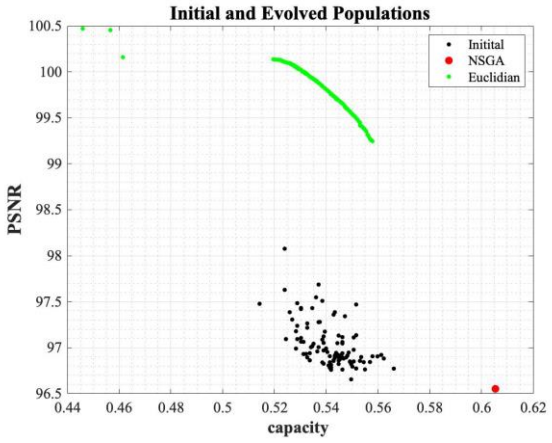


Figure 8. Distortion against capacity of the initial and evolved population tested on the T1-w MRI slice

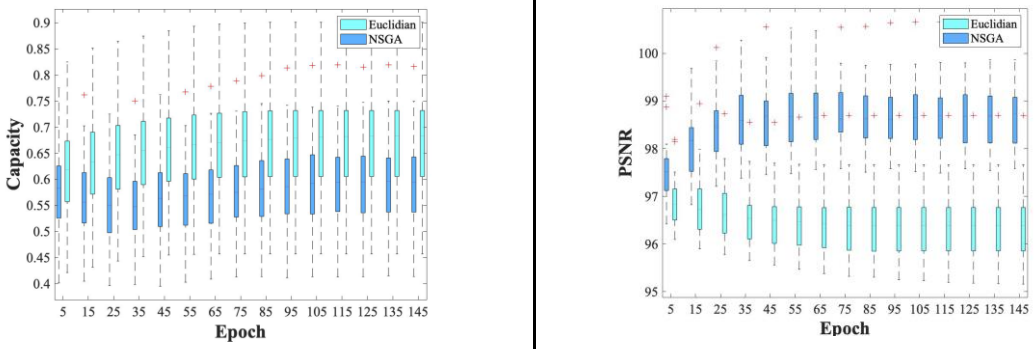
Table 2 provides a summary of the cluster characteristics of the whole dataset. There exist various numbers of clusters in the range of [115-640] with a high standard deviation. The zero bin numbers are given in the format of "a(s)/t", where *t* is the total number of zero bins that exist in the histogram. The number a indicates how many of the total *t* number of zero bins can be used for shifting. The number s denotes how many of the available zero bins can be paired with more than one cluster. The mean of the percentage of the ratio of the number *a* to the number *t* is 86.91 with a standard deviation of 0.06. In other words, around 13% of the total zero-bins are not considered to be used for shifting.

Table 1: Cluster Statistics.

	cluster count	clusterwidth (mean±std)	zero bins count (a(s)/t)*	maximum capacity
1	231	7.74±14.09	298(224)/335	2022
2	256	6.64±17.80	394(260)/439	2361
3	314	5.72±10.72	495(288)/588	2583
4	411	4.58±9.64	627(397)/744	2939
5	299	5.17±10.95	461(271)/496	2429
6	205	8.06±14.45	289(188)/326	1815
7	494	4.37±8.98	745(473)/886	3016
8	115	11.77±18.33	144(113)/181	1278
9	287	6.32±12.41	408(280)/463	2140
10	310	5.46±12.12	438(305)/479	2290
11	521	3.99±5.62	833(481)/967	3035
12	621	3.52±4.88	1020(564)/1317	3447
13	561	3.86±5.78	819(584)/872	2874
14	640	3.47±7.27	990(591)/1197	3400
15	178	8.59±12.90	266(160)/306	1725
16	118	12.76±23.80	145(122)/162	1130
17	222	7.25±16.49	300(233)/306	1771
18	360	5.46±7.05	516(368)/588	2260
19	327	5.82±8.79	490(312)/576	2390
20	198	8.17±11.40	257(216)/274	1468
21	383	5.27±11.06	604(379)/703	3202
22	325	5.60±8.31	464(320)/499	2192
23	291	6.21±12.86	405(274)/465	2313
24	151	10.56±23.20	206(163)/266	1777
25	458	4.50±7.03	704(446)/830	2937
26	132	11.20±35.07	182(137)/257	1887
27	249	7.11±18.39	344(240)/361	2063
28	348	5.54±9.66	475(366)/497	2268
29	308	5.91±11.93	435(288)/604	2667
30	262	6.86±10.26	351(279)/381	1914

∗ The number of zero bins are given in a(s)/t; t = total number of zero bins; a = available number of zero bins for shifting; s =among available bins, how many of them are shared by two clusters

Fig.9a and Fig.9b show the convergence characteristic curve of the capacity and PSNR values for training a GA model for 150 epochs. The populations rapidly converged to an approximate equilibrium almost 50 epochs later. The "multiple boxplot" MATLAB implementation of Ander Biguri (2021) was used to draw figures.



a) capacity vs epoch

b) PSNR vs epoch

Figure 9. *Convergence characteristics curves*

Furthermore, the comparison of the performance obtained using the solutions proposed by the individuals of the first population and the evolved population is given in Table 3 in terms of PSNR values. Left-tailed t-tests confirmed that the proposed method significantly improved the populations ($p<0.01$) when the NSA sorting method is applied. Only three cases of the experiment with Euclidian distance reported significant improvement in PSNR, nevertheless, less than the one with the NSA method.

C. Comparison with the case of using non-medical images

The histograms of the preprocessed non-medical test images Lena, Baboon and Pepper had a total of 308, 43 and 260 clusters, respectively. The Baboon image had fewer zero-bins than others with 85 zero-bins; on the other hand, only 65 percent of them could be included in the pairing procedure. Of the 424 zero bins, only 356 zero-bins were encoded in the GA model in the Lena image. After 150 epochs, the mean of the PSNR values of the individuals in the evolved populations is increased statistically significant compared to the initial populations for all three non-medical images, respectively, with a little improvement in the rate observed for the image Baboon.

D. Time complexity analysis

To compute T(n), the running time of proposed model for the given N_p number of individuals and N_e number of epochs as inputs into the training procedure, the products of the cost and times of each line in algorithm 1 are added together. The cost of each statement is a constant c_i , where i indicates the line number of the statement. The running time of a statement depends on how many times the statement is repeated rather than the cost of the statement [18]. Lines 2-3 are repeated n_p times, if n_p is expressed as p times n then the first loop of line 2-3 takes a time roughly proportional to n. There is a second loop between lines 6 and 12, which has an inner loop between the line 10 and the line 11. By considering n_e is a constant, the inner loop will be the main determining factor for the time complexity. Thus, based on the complexity analysis of the GA training , the GA model has a running time of O(n).

4. Conclusion

In this study, an evolutionary optimization algorithm for solving the rate-distortion trade-off in the HS-based multiple reversible data embedding algorithm is proposed. For this purpose, a specially designed genetic algorithm is proposed to determine the optimal pair of peak and zero bins for structural magnetic resonance imaging images. The chromosome encoding approach of the proposed GA model is inspired by the histogram characteristics of MR images. The proposed algorithm has been evaluated using a set of T1-weighted magnetic resonance images. The mean PSNR value of all individuals in the population after training with the proposed NSA-based GA procedure was statistically significantly increased up to 99.82. Experimental results show that the proposed algorithm yields higher quality embedded images without sacrificing embedding capacity in the field of medical imaging.

5. Acknowledgment

The data used in the preparation of this paper were obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers and clinical and neuropsychological.

References

[1] Zhicheng, N., Yun-qing, S., Ansari N, Su W. (2006) Reversible data hiding. IEEE Transactions On Circuits And Systems For Video Technology 16:354-362.

[2] Kiel, J. (2012). HIPAA and its effect on informatics. Comput Inform Nurs 30:1-5.

[3] Kurnaz, H., Konyar, M. Z., Sondaş, A. (2020). A new hybrid data hiding method based on near histograms, Eur J Sci Technol, 18:683-694.

[4] Vazhoramalayil, M., Vedhanayagam, M. (2021). A novel image scaling based reversible watermarking scheme for secure medical image transmission. Isa Transactions, 108:269-281.

[5] Wei-liang, T., Chia-ming, Y., Chin-chen, C. (2009). Reversible data hiding based on histogram modification of pixel differences. Ieee Transactions On Circuits And Systems For Video Technology, 19:906-910.

[6] Hwang, H., Kim, H., Vasiliy, S., Joo, S. (2010). Reversible watermarking method using optimal histogram pair shifting based on prediction and sorting. Journal of Transactions On Internet and Information Systems, 4:655-670.

[7] Kuo, W., Li, J., Wang, C., Wuu, L., Huang, Y. (2016). An improvement data hiding scheme based on formula fully exploiting modification directions and pixel value differencing method. IEEE Computer Society.

[8] Tian, J. (2003). Reversible data embedding using a difference expansion. IEEE Transactions On Circuits And Systems For Video Technology, 13:890-896.

[9] Malhotra, R., Singh, N., Singh, Y. (2011). Genetic algorithms: Concepts, design for optimization of process controllers. Comput Inf Sci, 4:39-5.

[10] Strossmayer, J. (2001). A Comparison of Several Heuristic Algorithms for Solving High Dimensional Optimization Problems 1.

[11] Wang, J., Ni, J., Zhang, X., Shi, Y. (2017). Rate and distortion optimization for reversible data hiding using multiple histogram shifting. Ieee Transactions On Cybernetics, 47:315-326.

[12] Wang, J., Ni, J. (10.1109/WIFS.2013.6707819). A GA optimization approach to HS based multiple reversible data hiding.

[13] Wang, J., Chen, X., Shi, Y. (2019). Unconstraint optimal selection of side information for histogram shifting based reversible data hiding. Ieee Access 35564-35578.

[14] Kuo-liang, C., Yong-huai, H., Wei-ning, Y., Yu-chiao, H., Chyou-hwa, C. (2009). Capacity maximization for reversible data hiding based on dynamic programming approach. Applied Mathematics And Computation, 208:284-292.

[15] Qi, W., Li, X., Zhang, T., Guo, Z. Optimal Reversible Data Hiding Scheme Based on Multiple Histograms Modification. Ieee Transactions On Circuits And Systems For Video Technology, 2300-2312.

[16] Wansapura, J., S., Dunn, R., Ball, W. (1999). NMR relaxation times in the human brain at 3.0 tesla. Journal Of Magnetic Resonance Imaging

[17] Hazlewood, C., Yamanashi, W., Rangel, R., Todd, L. (1982). In vivo NMR imaging and T1 measurements of water protons in the human brain. Magnetic Resonance Imaging, 1:3-10.

[18] Cormen, T.,H. (2009). Introduction to algorithms. MIT press, 658- 662.