



HAL
open science

A Comparative Study of Two State-of-the-Art Feature Selection Algorithms for Texture-Based Pixel-Labeling Task of Ancient Documents

Maroua Mehri, Ramzi Chaieb, Karim Kalti, Pierre Héroux, Rémy Mullot, Najoua Essoukri Ben Amara

► **To cite this version:**

Maroua Mehri, Ramzi Chaieb, Karim Kalti, Pierre Héroux, Rémy Mullot, et al.. A Comparative Study of Two State-of-the-Art Feature Selection Algorithms for Texture-Based Pixel-Labeling Task of Ancient Documents. *Journal of Imaging*, 2018, 4 (8), pp.97. 10.3390/jimaging4080097. hal-04091054

HAL Id: hal-04091054

<https://hal.science/hal-04091054v1>

Submitted on 19 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

A Comparative Study of Two State-of-the-art Feature Selection Algorithms for Texture-based Pixel-labeling Task of Ancient Documents

Maroua Mehri ^{1,2} , Ramzi Chaieb ¹, Karim Kalti ¹, Pierre Héroux ², Rémy Mullot ³
and Najoua Essoukri Ben Amara ¹

¹ LATIS Laboratory, Sousse University, National Engineering School of Sousse, 4023, Sousse Erriadh, Tunisia; maroua.mehri@gmail.com; ramzi.chaieb@hotmail.com; karim.kalti@gmail.com; benamaranajwa@gmail.com

² LITIS Laboratory, Normandie University, Avenue de l'Université, 76800, Saint-Etienne-du-Rouvray, France; pierre.heroux@univ-rouen.fr

³ L3i Laboratory, University of La Rochelle, Avenue Michel Crépeau, 17042, La Rochelle, France; remy.mullot@univ-lr.fr

* Correspondence: maroua.mehri@gmail.com

Academic Editor: name

Version July 21, 2018 submitted to J. Imaging

Abstract: Recently, texture features have been widely used for historical document image analysis. However, few studies have focused exclusively on feature selection algorithms for historical document image analysis. Indeed, an important need has emerged to use a feature selection algorithm in data mining and machine learning tasks, since it helps to reduce the data dimensionality and to increase the algorithm performance such as a pixel classification algorithm. Therefore, in this paper we propose a comparative study of two conventional feature selection algorithms, genetic algorithm and ReliefF algorithm, using a classical pixel-labeling scheme based on analyzing and selecting texture features. The two assessed feature selection algorithms in this study have been applied on a training set of the HBR dataset in order to deduce the most selected texture features of each analyzed texture-based feature set. The evaluated feature sets in this study consist of numerous state-of-the-art texture features (Tamura, local binary patterns, gray-level run-length matrix, auto-correlation function, gray-level co-occurrence matrix, Gabor filters, 3-level Haar wavelet transform, 3-level wavelet transform using 3-tap Daubechies filter and 3-level wavelet transform using 4-tap Daubechies filter). In our experiments, a public corpus of historical document images provided in the context of the historical book recognition contest (*HBR2013 dataset*) has been used. Qualitative and numerical experiments are given in this study in order to provide a set of comprehensive guidelines on the strengths and the weaknesses of each assessed feature selection algorithm according to the used texture feature set.

Keywords: benchmarking; texture; feature selection; pixel-labeling; ancient document images

1. Introduction

Providing reliable computer-based access and analysis of cultural heritage documents has been flagged as a very important need for the library and the information science community, spanning educationalists, students, practitioners, researchers in book history, computer scientists, historians, librarians, end-users and decision makers. More specifically, there is a consistent and clear need for robust and accurate document image analysis (DIA) methods that deal with the idiosyncrasies of historical document images [1,2]. Indeed, historical DIA remains an open issue due to the particularities

27 of historical documents, such as the superimposition of information layers (e.g. stamps, handwritten
28 notes, noise, back-to-front interference, page skew) and the variability of their contents and/or layouts.
29 Moreover, analyzing historical document images and characterizing their layouts and contents under
30 significant degradation levels and different noise types and with no *a priori* knowledge about the
31 layout, content, typography, font styles, scanning resolution or DI size, *etc.* is not a straightforward
32 task. Therefore, researchers specialized in historical DIA keep proposing novel reliable approaches
33 and rigorous techniques for historical DIA, segmentation and characterization. Recently, there has
34 been increasing interest in using deep architectures for solving various sub-fields and tasks related to
35 the issues surrounding computer vision and pattern recognition and particularly document image
36 analysis and handwritten text recognition. For instance, deep neural networks have developed for
37 feature learning [24] and document layout and content analysis [4,5]. For instance, Chen *et al.* [5]
38 proposed a pixel-labeling approach for handwritten historical document images segmentation based
39 on using a convolutional neural network (CNN). Calvo-Zaragoza *et al.* [4] presented a CNN-based
40 method for automatic document processing of music score images. Wei *et al.* [24] proposed a layout
41 analysis method of historical document images using the sequential forward selection algorithm and
42 the autoencoder technique as a deep neural network for feature selection and learning. Nevertheless,
43 these methods based on deep architectures are hindered by many issues related to the computational
44 cost in terms of memory consumption, processing time and computational complexity on the one
45 hand, and the need for large datasets.

46 In the literature, the methods used for DIA have been classified into two categories: texture
47 and non-texture-based [13]. Kise [10] stated that the most relevant DIA methods used to analyze
48 documents with unconstrained layouts and overlapping layers are based on texture features. It
49 has been demonstrated that the text/graphic region separation task can be performed efficiently by
50 using a texture-based method. On the other hand, the textual regions with different fonts can be
51 segmented using texture features which are often used for text font characterization. A text font is
52 mainly characterized by its weight, style, condensation, width, slant, italicization, ornamentation, and
53 designer or foundry [20].

54 However, using a texture-based method has quite high computational complexity since it often
55 involves a large number of features. Indeed, two criteria can be identified when using a texture-based
56 method: object to be analyzed (*i.e.* foreground or background) and primitive of analysis (*i.e.* pixels,
57 superpixels, connected components, *etc.*). These two criteria entail large volumes of data to be processed
58 when using a texture-based method. Moreover, the processing time of a texture-based method depends
59 entirely on the image size and resolution due to the use of a primitive-based computation. However,
60 there is awareness that maybe there are redundant and non-relevant indices when extracting and
61 analyzing texture features which may reduce the performance of a texture-based algorithm. Feature
62 selection meets this real need by selecting relevant features and by removing redundant ones in order
63 to reduce the data dimensionality, to improve the quality of the feature set and to increase the algorithm
64 performance, such as a texture-based pixel-labeling algorithm.

65 Thus, in this paper a comparative study of two conventional feature selection algorithms, genetic
66 algorithm (GA) and ReliefF algorithm (RA), is proposed in order to provide a set of comprehensive
67 guidelines on the strengths and the weaknesses of each assessed feature selection algorithm according
68 to the used texture feature set. The texture-based feature sets which have been compared and evaluated
69 in this study have been derived from the Tamura, local binary patterns (LBP), gray-level run-length
70 matrix (GLRLM), auto-correlation, gray-level co-occurrence matrix (GLCM), Gabor filters and three
71 wavelet-based approaches: 3-level Haar wavelet transform (Haar), 3-level wavelet transform using
72 3-tap Daubechies filter (Db3) and 3-level wavelet transform using 4-tap Daubechies filter (Db4).

73 In our comparative study, a public corpus of historical document images (called the *HBR2013*
74 *dataset*) which was provided by the pattern recognition and image analysis research lab (PRIma)¹
75 has been used [1,2]. The *HBR2013 dataset* has been proposed in the context of the historical book
76 recognition (HBR) contest held in conjunction with the ICDAR conference (2011 and 2013). The
77 *HBR2013 dataset* is a subset of the IMPACT dataset², representing key holdings of major European
78 libraries and consisting of printed documents of various types (e.g. books, newspapers, journals, legal
79 documents), in 25 languages from the 17th century to the early 20th century. It contains a large diversity
80 of historical document contents (variety of layouts and contents). The *HBR2013 dataset* presents many
81 particularities and challenges which motivates us to conduct our thorough study on it.

82 The remainder of this article is organized as follows. Sections 2 and 3 review firstly the
83 texture-based methods and feature selection algorithms proposed in the literature, respectively, with a
84 particular focus on those related to historical DIA. A brief report of the different texture-based feature
85 sets and feature selection algorithms evaluated in this study is also given. Section 4 describes the
86 experimental protocol by firstly presenting the main phases of the proposed pixel-labeling scheme
87 used for analyzing and comparing the performance of each texture feature set according to the use of a
88 full texture feature set, the use of a subset of texture features selected by means of the GA, and the use
89 of a subset of texture features selected by means of the RA (*cf.* Section 4.1). Secondly, the experimental
90 corpus and the defined ground truth used in our experiments are detailed in 4.2. Then, qualitative
91 results are given to demonstrate the performance of each texture-based feature set according to the
92 use or not of a feature selection algorithm (*cf.* Section 4.4). Afterwards, we discuss quantitatively
93 the obtained performance of the texture feature analysis experiments (*cf.* Section 4.4). Finally, our
94 conclusions and future work are presented in Section 5.

95 2. Texture features

96 Recently, many DIA issues have been focused on using texture-based approaches for segmentation
97 and classification tasks [13]. Indeed, the use of texture analysis techniques for historical document
98 images has become an appropriate choice, since it has been shown that texture-based approaches
99 work effectively with no *a priori* knowledge about the layout, content, typography, font and graphic
100 styles, scanning resolution, document image size, *etc.* Moreover, the use of a texture-based approach
101 has been shown to be effective with skewed and degraded images. Therefore, the interest in using a
102 texture-based method for historical DIA is continuously increasing [12].

103 In the literature, based on extracting and analyzing texture features a texture-based method has
104 been usually used to partition the analyzed image into regions. The obtained regions have similar
105 properties and characteristics with respect to the extracted texture features [3]. Thus, this study is
106 based on the two following assumptions: text regions have different texture features from non-text
107 ones and textual regions with different fonts are also distinguishable [13].

108 Relatively a limited number of comparative studies address the problem of presenting quantitative
109 comparisons of texture-based algorithms, although it is commonly agreed that texture analysis plays a
110 fundamental role for DIA [6]. Visual or qualitative results of seven texture-based methods (run-lengths,
111 multi-channel Gabor filters, texture co-occurrence spectrum, white tiles, texture masks, structured
112 wavelet packet analysis and laws masks) have been reported in [13]. Mehri *et al.* [12] presented
113 a benchmarking of the most classical and widely used texture-based feature sets which had been
114 conducted using a classical texture-based pixel-labeling scheme on a corpus of historical document
115 images. This comparative study has been carried out for selecting the most relevant texture feature set
116 based on the best trade-off between the best performance and the lowest computation time.

1 <http://www.primaresearch.org>

2 <http://www.primaresearch.org/datasets>

117 Therefore, the texture-based features which are compared and evaluated in this article have been
118 derived from the Tamura, LBP, GLRLM, auto-correlation, GLCM, Gabor filters and three wavelet-based
119 approaches: Haar, Db3 and Db4.

120 3. Feature selection algorithms

121 Using a texture-based method often involves a large number of texture features in
122 high-dimensional spaces to be analyzed. Indeed, each analyzed image will be described by a set
123 of multi-dimensional texture-based feature vectors. This will induce greater computational cost and
124 occupy a lot of storage space since a large and complex feature space has been generated. Moreover,
125 it is worth noting that the smaller the dimension of the analyzed texture-based space, the easier it
126 will be to deal with the specified task. Besides, if the number of dimensions becomes very large, this
127 will make it more difficult to compute data similarity and perform data mining tasks. Indeed, the
128 data similarity is sensitive to the number of dimensions (curse of dimensionality) since it is based on
129 computing distance between vectors (*i.e.* the higher the number of dimensions, the higher the values
130 of distance between vectors and the more difficult it will be to group data).

131 Based on these findings, redundant or even irrelevant features may affect the learning process
132 and consequently reduce the pixel classification accuracy in the case of our work. For instance, Journet
133 *et al.* [9] extracted three auto-correlation features and two frequency descriptors by using a multi-scale
134 analysis for classifying pixels into text, graphics and background in historical document images. Then,
135 they proposed to reduce the dimension space of the extracted features using the principal component
136 analysis technique. They demonstrated that only 78% of the extracted features are relevant. In order to
137 classify pixels from historical document images into four classes (periphery, background, text block,
138 and decoration), Wei *et al.* [23] used the convolutional auto-encoder features and concluded that more
139 than 80% of the analyzed features are redundant or irrelevant.

140 Therefore, a feature selection phase is often required to avoid these problems by selecting the
141 most relevant features and remove redundant ones from the original large set of texture-based features
142 [25]. Sequential forward selection, sequential backward selection, tabu search, genetic algorithm and
143 ReliefF algorithm are the most well-known and widely used feature selection algorithms [26]. A
144 feature selection algorithm is based on using a search technique to evaluate different proposals of
145 feature subsets by means of an evaluation measure in order to determine the one that has the best
146 performance [8].

147 Figure 1 depicts the common key steps of a feature selection process. The general procedure for
148 feature selection starts by creating a candidate feature subset for evaluation. Each candidate subset is
149 evaluated by using an evaluation criterion to measure the quality of the selected features. The process
150 of subset generation and evaluation is repeated until a predefined stopping criterion is satisfied. The
151 feature selection process ends by outputting the selected subset of features to a validation procedure.

152 Few researchers have addressed feature selection issues for historical DIA. For instance, Tao *et al.*
153 [19] proposed a feature selection algorithm based on using the LBP operator and dimension reduction
154 technique for Chinese character font categorization. A hybrid feature selection method was proposed
155 by Wei *et al.* [22] for historical DIA. The proposed feature selection method was based on using an
156 adapted greedy forward selection method and the genetic selection algorithm in a cascading way
157 to select different kinds of features including color, gradient, and LBP. By comparing their method
158 with four conventional feature selection methods (genetic selection, linear forward Selection, best
159 first forward selection and best first backward selection), Wei *et al.* [22] concluded that their method
160 selected significantly fewer features and provided lower error rates. They also concluded that the
161 most discriminative features for layout analysis of documents of diverse nature are the LBP ones. In
162 our paper, we have focused on the multi-scale texture analysis of historical document images using
163 nine texture feature sets (Tamura, LBP, GLRLM, auto-correlation, GLCM, Gabor filters, Haar, Db3 and
164 Db4). However, Wei *et al.* [22] investigated three main sets of texture features (color, gradient and LBP
165 features) without using a multi-scale analysis. They combined all these features in a 204-dimensional

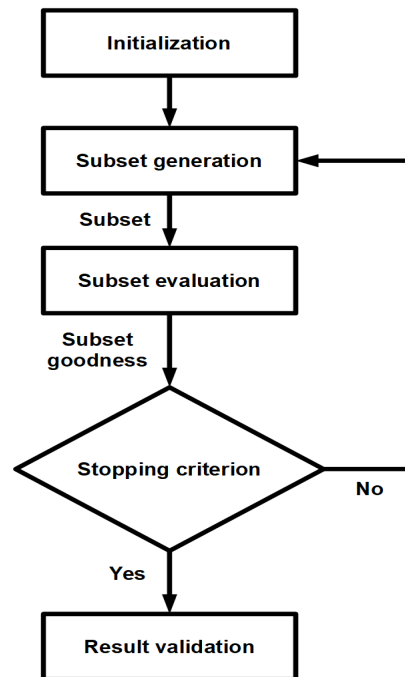


Figure 1. Common key steps of a feature selection process.

166 feature vector. Furthermore, we have investigated separately the two feature selection algorithms
 167 (genetic and ReliefF algorithms) on each texture feature set. However, a cascading feature selection
 168 method (a cascade of an adapted forward selection and a genetic selection algorithms) was proposed
 169 in [22]. Besides, comparing to [22] we have used more images (60 images) during the training phase.

170 To the best of our knowledge, there is no comparative study that has been carried out to investigate
 171 jointly the most well-known texture-based feature sets and widely used feature selection algorithms
 172 for historical DIA. Therefore, we propose in this article to evaluate the use of two conventional feature
 173 selection algorithms, genetic algorithm and ReliefF algorithm, in order to select an optimal subset of
 174 each texture-based feature set for pixel-labeling task in ancient document images.

175 3.1. Genetic algorithm

176 The genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. First,
 177 a population of chromosomes which encodes candidate solutions is created. A chromosome is a string
 178 of bits (1 and 0 indicate whether a feature is selected or not, respectively) whose size corresponds to the
 179 number of features. Then, the solutions are evolved by applying genetic operators such as crossover
 180 and mutation to find the best solution based on a predefined fitness function. Commonly, the GA
 181 terminates when either a maximum number of generations has been produced or a satisfactory fitness
 182 level has been reached for the population [7]. Algorithm 1 details the different parameters used in the
 183 GA. More details were given in [14] with a thorough description of the different parameters used in
 184 the GA.

Figure 2 presents a flowchart summarizing the fundamental steps of the GA used in this study. The GA starts by creating an initial population of randomly generated individuals using the following formula:

$$P = \text{round}((L - 1) \times \text{rand}(DF, 200 \times DF)) + 1 \quad (1)$$

185 where L and DF represent the number of input features and the desired number of selected features,
 186 respectively. In the GA experiments, DF is set to $L/2$.

Algorithm 1 Basic genetic algorithm [7]

Input: Crossover probability (P_{co})
 Mutation probability (P_{mut})
 Population size (L-chromosomes- or classifier- by N-bits)
 Criteria function ($Fit()$)
 Fitness threshold (θ)

Output: Set of highest fitness chromosomes (best classifier)

```

1: repeat
2:   Determine the fitness of each chromosome:  $Fit(i), i = 1, \dots, L$ 
3:   Rank the chromosomes
4:   repeat
5:     Select two chromosomes with highest score
6:     if  $Rand[0, 1] < P_{co}$  then
7:       Crossover the pair at a randomly chosen bit
8:     else
9:       Change each bit with the probability  $P_{mut}$ 
10:    Remove the parent chromosomes
11:   until  $N$  offspring have been created
12: until Any chromosome's score  $Fit()$  exceeds  $\theta$ 
13: return Highest fitness chromosome (best classifier)

```

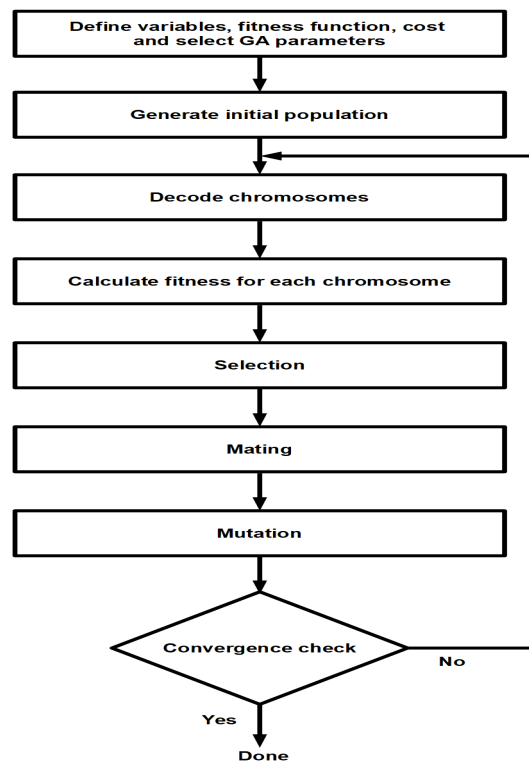


Figure 2. Flowchart of the GA.

In each generation, a proportion of the existing population is selected to breed a new generation. Each selected individual solution is evaluated on the basis of its overall fitness. In the GA experiments, a fitness function based on the principle of Minimum Redundancy Maximum Relevance (*mRMR*)

is used [14]. The key idea of *mRMR* is to select the set S with m features $\{x_i\}$ that satisfies the maximization problem:

$$\max \Phi_i(D, R); \Phi(D, R) = D - R \quad (2)$$

187 where D and R represent the max-relevance and min-redundancy, respectively. D and R are defined as
188 follows:

$$D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, y) \quad (3)$$

$$R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \quad (4)$$

where $I(x_i, y)$ and $I(x_i, x_j)$ represent the mutual information, which is the quantity that measures the mutual dependence of the two random variables and is calculated as follows:

$$I(x, y) = H(x) + H(y) - H(x, y) \quad (5)$$

189 where $H(\cdot)$ is the entropy.

190 3.2. ReliefF algorithm

191 The ReliefF algorithm (RA) is one of the most famous feature weighting methods. It assigns a
192 weight to each feature, and the features values over a particular threshold are selected. The key idea of
193 the RA is to select features randomly, and then based on nearest neighbors the relevance of features
194 according to how well their values distinguish among the instances of the same and different classes
195 that are near to each other is estimated [17]. The bigger the weight value, the better the feature is.
196 Algorithm 2 gives a more detailed description of the process of the RA method. More details were
197 given in [18] with a thorough description of the key steps of the investigated RA.

Algorithm 2 ReliefF algorithm [18]

Input: For each training instance:

Vector of attribute values ($A_i, i = 1, \dots, a$)

Class value (C)

Output: Vector W of the estimations of the qualities of attributes

- 1: Set all weights $W[A] := 0.0$
- 2: **for** $i := 1$ to m **do**
- 3: Randomly select an instance R_i
- 4: Find k nearest hits H_j
- 5: **for each** class $C \neq \text{class}(R_i)$ **do**
- 6: From class C find k nearest misses $M_j(C)$
- 7: **for** $A := 1$ to a **do**

$$W[A] := W[A] - \sum_{j=1}^k \frac{\text{diff}(A, R_i, H_j)}{m \times k} + \sum_{C \neq \text{class}(R_i)} \frac{\frac{P(C)}{1 - P(\text{class}(R_i))} \sum_{j=1}^k \text{diff}(A, R_i, M_j(C))}{m \times k}$$

where m is a user-defined parameter. $\text{diff}(A, I_1, I_2)$ is a function that computes the difference between the values of the attribute A for two instances I_1 and I_2 . $P(\cdot)$ denotes the prior probability.

198 4. Evaluation and results

199 In this section, a brief description of the main phases of the pixel-labeling scheme used for
 200 analyzing and selecting texture features is presented. Then, qualitative results are given to demonstrate
 201 the performance of each texture-based feature set according to the use or not of a feature selection
 202 algorithm. Subsequently, the performance of each texture feature set according to the use of a full
 203 texture feature set, the use of a subset of texture features selected by means of the GA, and the use of a
 204 subset of texture features selected by means of the RA is discussed after describing our experimental
 205 corpus and its associated ground truth, and presenting the used accuracy metrics for performance
 206 evaluation.

207 4.1. Pixel-labeling scheme

208 In order to investigate the importance of using a feature selection algorithm for historical DIA,
 209 a generic and standard framework that ensures a fair analysis and comparison of performance is
 210 required. The proposed framework is presented in this study as a pixel-labeling scheme based on
 211 analyzing and selecting texture features. It aims at analyzing and comparing of the performance of
 212 each texture feature set according to the use of a full texture feature set, the use of a subset of texture
 213 features selected by means of the GA, and the use of a subset of texture features selected by means of
 214 the RA.

215 The main goal of the proposed pixel-labeling consists of structuring the texture feature space
 216 within a clustering technique in order to group pixels sharing similar characteristics. The proposed
 217 pixel-labeling scheme forms the basis of a classical layout analysis approach and cornerstone of
 218 different DIA tasks related to segmentation, analysis, classification and recognition of historical
 219 document images, *etc.* The pixel-labeling scheme used in our experiments to analyze and select texture
 220 features is illustrated in Figure 3.

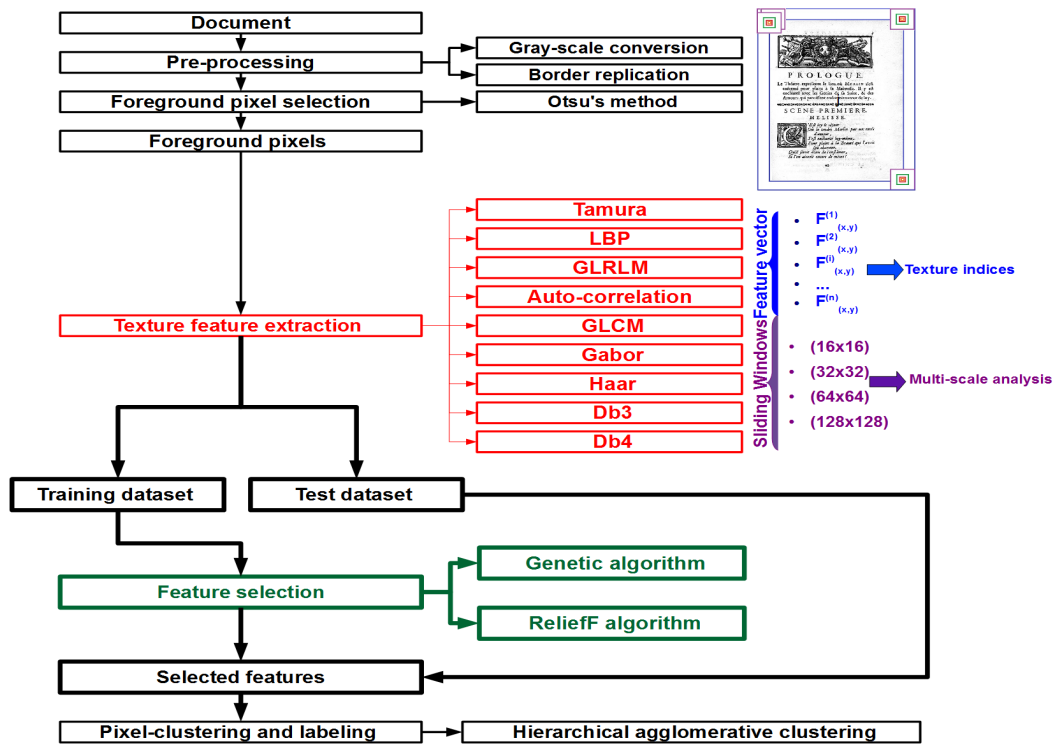


Figure 3. Pixel-labeling scheme based on analyzing and selecting texture features.

221 First of all, each historical document image of our experimental corpus is fed as input of our
222 proposed pixel-labeling scheme. Then, texture feature have been extracted only from the foreground
223 pixels of gray-scale images without using any binarization step. By using analysis windows of varying
224 sizes (*i.e.* a pixel-wise technique), the texture feature extraction step is performed in order to adopt a
225 multi-resolution/multi-scale approach. By using a multi-scale approach, more reliable information
226 can be obtained and region boundaries can be identified more accurately since textural characteristics
227 can be perceived differently at varying scales. A border replication step is applied on each image in
228 order to deal with foreground pixels located at image borders when computing texture features.

229 Then, all extracted features have been used as input for both the GA and the RA individually.
230 Two separate datasets, namely, the training dataset (60%) and the testing dataset (40%) that our
231 experimental corpus comprises have been used separately in our experiments. A learning phase
232 is introduced in the proposed pixel-labeling scheme that the most selected texture features will be
233 identified according to the textural characteristics of a 60% of document images selected randomly
234 from the *HBR2013 dataset*. For each document image in the training dataset, only 50% of all the features
235 have been selected when performing separately the GA and the RA iterations. Afterwards, the subset
236 of the most selected texture features used on evaluating the testing dataset is deduced based on the
237 following heuristic: a texture feature would be counted among the subset of the most selected texture
238 features by using a feature selection algorithm, if it was chosen by over half the images of the training
239 dataset.

240 Given the results of the most selected texture features from the training dataset, an unsupervised
241 clustering step is afterwards performed based on analyzing the subset of the most selected texture
242 features extracted from the foreground pixels of the testing dataset. The clustering step is performed
243 by using the hierarchical ascendant classification (HAC) algorithm and by setting the number of
244 homogeneous and similar content regions (k) equal to the one defined in the ground truth in order
245 to avoid inconsistencies and bias in assessments caused by estimating automatically the number of
246 homogeneous and similar content regions and subsequently to ensure an objective understanding of
247 the behavior of the evaluated texture feature sets and feature selection algorithms. The HAC algorithm
248 is performed on the computed texture features without taking into account the spatial coordinates.
249 The HAC algorithm process consists of successively merging pairs of existing clusters where at each
250 cluster grouping step, the choice of cluster pairs depends on the smallest distance (*i.e.* clusters are
251 grouped if the intra-cluster inertia is minimal). This linkage between clusters is performed using the
252 Ward criterion along with the weighted Euclidean distance [21].

253 By using the HAC algorithm the obtained texture-based feature vector sets are partitioned into k
254 compact and well-separated clusters in the multi-dimensional feature space, producing a pixel-labeled
255 image as output. Since the used classifier process in the pixel-labeling scheme is unsupervised, the
256 colors attributed to the different document image contents (text or graphics) may differ from one
257 document image to another.

258 4.2. Corpus and preparation of ground truth

259 In our experiments, a public corpus of historical document images provided in the context of the
260 HBR contest (*HBR2013 dataset*) has been used. The *HBR2013 dataset* contains 100 binary, gray-scale or
261 color historical document images which were digitized at 150/300 dpi. Table 1 details the *HBR2013*
262 *dataset* characteristics. Figure 4 illustrates samples of pages of the *HBR2013 dataset*.

263 To analyze the performance of each texture-based feature set according to the use or not of
264 a feature selection algorithm in the proposed pixel-labeling scheme, a pixel-based ground truth is
265 required. For this purpose, the ground truthing environment for document images (GEDI)³ has been
266 used in our experiments.

³ <http://gedigroundtruth.sourceforge.net/>

Table 1. Composition of the *HBR2013 dataset*.

Content	Number of pages	Number of fonts	Graphics
Only one font (cf. Figure 4[a])	3	1	No
Only two fonts (cf. Figure 4[b])	17	2	No
Graphics and text with two different fonts (cf. Figure 4[c])	9	2	Yes
Only three fonts (cf. Figure 4[d])	20	3	No
Graphics and text with three different fonts (cf. Figure 4[e])	6	3	Yes
Only four fonts (cf. Figure 4[f])	11	4	No
Graphics and text with four different fonts (cf. Figure 4[g])	15	4	Yes
Only five fonts (cf. Figure 4[h])	5	5	No
Graphics and text with five different fonts (cf. Figure 4[i])	14	5	Yes

267 Our ground truth has been manually outlined by labeling spatial boundaries of regions annotating
 268 the textual and graphical contents. Figure 5 illustrates few examples of the defined ground truth.
 269 Different labels for regions with different fonts have been also annotated for evaluating the performance
 270 of texture feature to separate various text fonts. Then, to provide a pixel-accurate representation of the
 271 analyzed images of the *HBR2013 dataset*, each selected foreground pixel is annotated according to the
 272 label of the region to which it belongs.

273 Analyzing the nine sets of texture descriptors and two feature selection algorithms using the
 274 *HBR2013 dataset* gives a total of 1800 analyzed images (100 images \times 9 different texture-based
 275 approaches \times 2 different feature selection algorithms).

276 4.3. Qualitative results

277 A visual comparison of the resulting images of historical document examples of the training and
 278 testing datasets of the *HBR2013 dataset* using the proposed pixel-labeling scheme is discussed in this
 279 section.

280 Figure 6 depicts the resulting images of a historical document example of the “Three fonts and
 281 graphics” category of the training dataset of the *HBR2013 dataset*, while Figure 7 illustrates the resulting
 282 images of a historical document example of the “Three fonts and graphics” category of the testing
 283 dataset of the *HBR2013 dataset*. The number of class labels in the resulting images is equal to 4. Since
 284 the pixel-labeling task is unsupervised, the colors attributed to text or graphics may differ from one
 285 document to another.

286 From the series of the resulting images given in the two Figures 6 and 7, we see that the obtained
 287 results are slightly astounding. For instance, the best pixel-labeling results are given by analyzing
 288 the selected Gabor features by means of the GA when the analyzed document belongs to the training
 289 dataset (*i.e.* graphical regions in blue color are more homogeneous), which is not the case when
 290 the analyzed document belongs to the testing dataset (cf. Figure 6[s]). This can be justified by the
 291 particularities of the *HBR2013 dataset* (strong heterogeneity, with differences in layout, typography,
 292 illustration style, complex layouts and historical spelling variants, *etc.*) since it consists of printed
 293 documents of various types (e.g. books, newspapers, journals, legal documents). It represents a wide
 294 variety of layouts that reflect several particularities of historical document images. This points out that
 295 applying a global selection on the *HBR2013 dataset* is not quite relevant that it is necessary to train on
 296 documents having similar characteristics in terms of the layout structure and/or typographic/graphical
 297 properties of the historical document image content. The quality of the pixel-labeling task will be more
 298 convincing if we use a feature selection algorithm on documents having some similarities of document

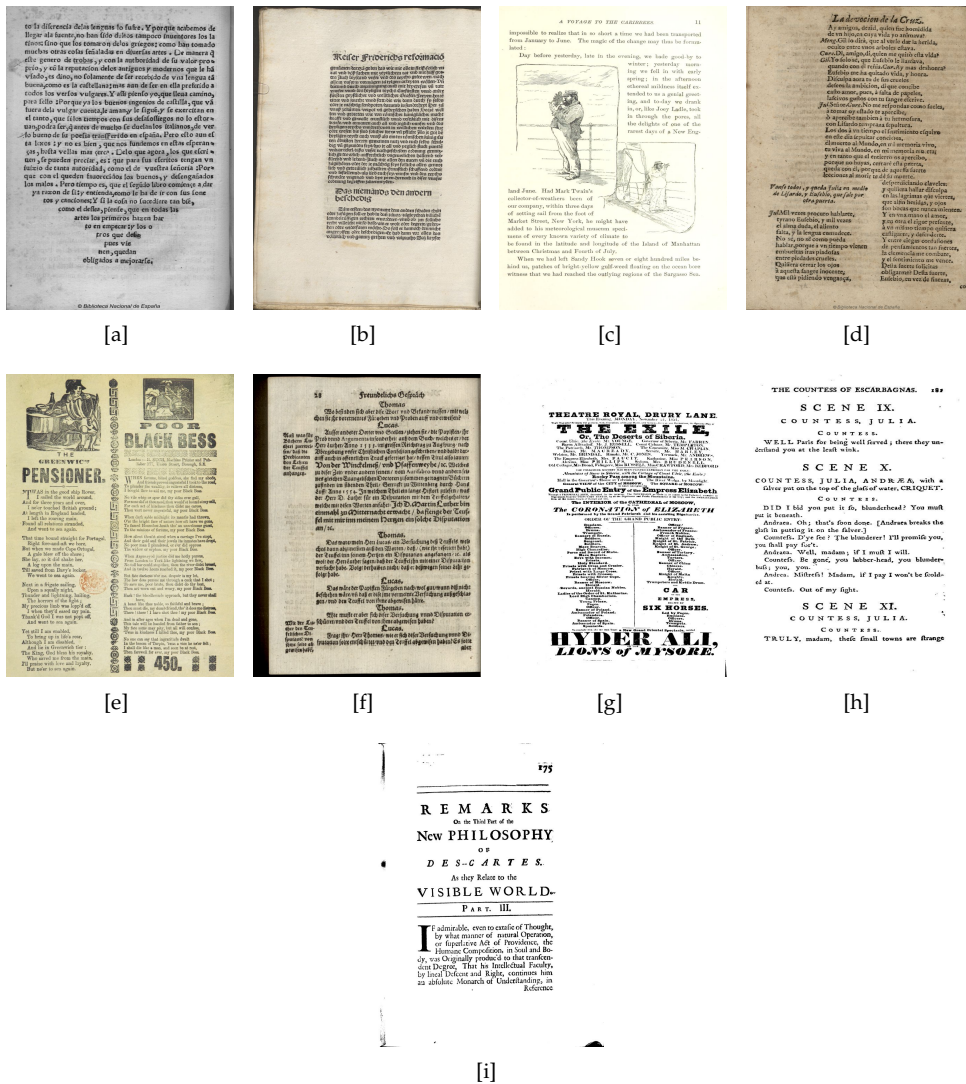


Figure 4. Historical document image examples of the *HBR2013* dataset. Figures 4[a], 4[b], 4[c], 4[d], 4[e], 4[f], 4[g], 4[h], and 4[i] illustrate examples of historical document images of the *HBR2013* dataset containing only two fonts, two fonts and graphics, only three fonts, three fonts and graphics, only four fonts, four fonts and graphics, only five fonts and five fonts and graphics, respectively.

299 content type (some similarities of document content type can be deduced from many book pages since
 300 a document content type can be repeated on many pages of the same book).

301 By comparing the visual results of a document belonging to the testing dataset, we note a drop in
 302 performance in terms of homogeneity when the analyzed features are given by selecting the LBP and
 303 Gabor features by means of the GA (cf. Figures 7[g] and 7[s]) and by means of the RA (cf. Figures 7[h]
 304 and 7[t]). In Figure 7[s], we show that some foreground pixels characterizing a textual content (cyan)
 305 has been labeled as graphical one (green and blue), while in Figure 7[t] we see that some foreground
 306 pixels characterizing a graphical content (red, green, and blue) has been labeled as textual one (cyan).

307 We also show that the results have significantly improved when using in the proposed
 308 pixel-labeling scheme the Tamura features selected using the RA on documents of the training and
 309 testing datasets (cf. Figures 6[e] and 7[e]). We observe that when using the selected GLRLM features
 310 by means of the GA and RA algorithms on a document of the testing dataset, the pixel-labeling
 311 quality has improved considerably (cf. Figures 7[j] and 7[k]), unlike when using the selected
 312 auto-correlation features (cf. Figures 7[m] and 7[n]). The pixel-labeling results given by analyzing the

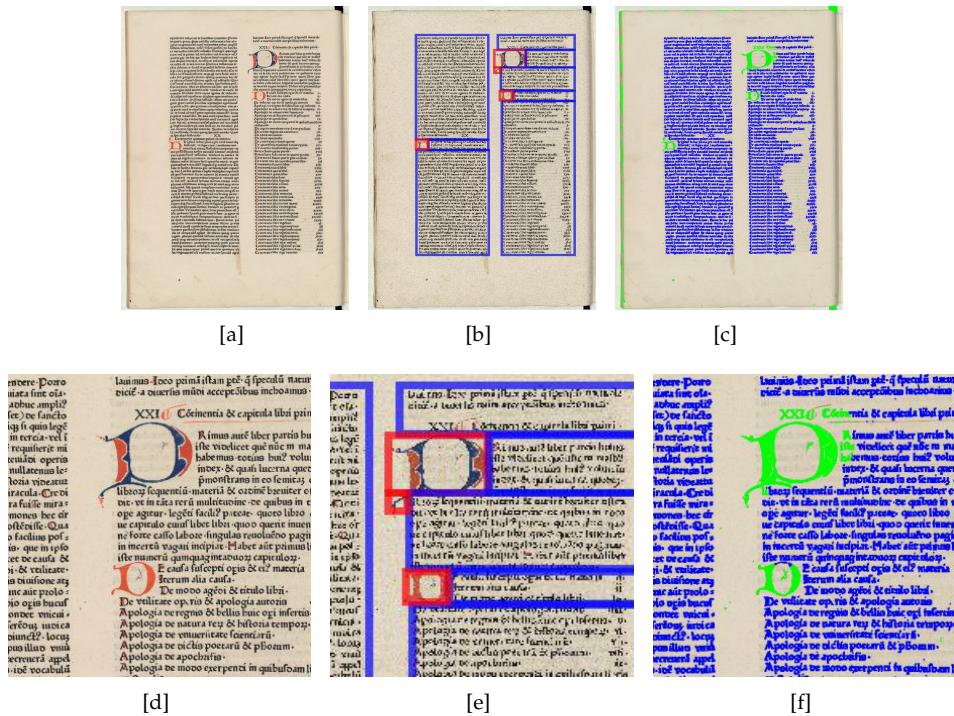


Figure 5. Example of the defined ground truth and obtained pixel-labeling result. Figures 5[a] and 5[b] illustrate an original historical document image and its associated ground truth, respectively. Figure 5[c] shows the final result of the pixel-labeling task by analyzing the Gabor features. Figures 5[d], 5[e], and 5[f] illustrate zoomed regions of Figures 5[a], 5[b], and 5[c], respectively.

313 full auto-correlation feature set (*cf.* Figure 7[l]) on the proposed pixel-labeling scheme on a document
 314 of the testing dataset are relatively similar to those based on selecting auto-correlation features by
 315 means of a feature selection algorithm (*cf.* Figures 7[m] and 7[n]).

316 We see that the Gabor and Db4 features give the best results in terms of the homogeneity of the
 317 textual region content when using in the proposed pixel-labeling scheme the full texture feature set (*cf.*
 318 Figures 7[r] and 7[aa]) and the texture features selected using the RA (*cf.* Figures 7[t] and 7[ac]) on
 319 a historical document example of the testing dataset. We also note that in the case of using the full
 320 Gabor and Db4 feature sets, the Gabor and Db4 features selected using the RA, the textual regions
 321 with different sizes and fonts have not been separated properly and particularly when the documents
 322 also contain graphics (more than one cluster is assigned for graphical regions by discriminating many
 323 orientations that are present to different extents in graphical regions). This confirms that the Gabor
 324 and Db4 features characterize specifically the main orientation of a texture. A suitable alternative is to
 325 use a recursive clustering method in order to ensure the distinction between distinct text fonts and
 326 various graphic types when the documents under consideration are complex and contain graphics and
 327 various kinds of fonts.

328 4.4. Benchmarking and performance evaluation

329 The dimensionality and performance evaluation of each texture-based feature set in the following
 330 three cases: with full texture feature set, with texture features selected using the GA, and with texture
 331 features selected using the RA, using the proposed pixel-labeling scheme on the *HBR2013 dataset* are
 332 presented in Table 2.

The Gabor and GLRLM signatures have the largest dimensions equal to 192 and 176, respectively, while the Tamura and auto-correlation signatures have the smallest dimensions equal to 16 and 20, respectively. By applying the GA and RA on a document of the training dataset, the number of features

has been reduced by half. We note that the number of features has been significantly reduced. The reduction ratio (RD) is computed using the following equation:

$$RD = 1 - \frac{N'_f}{N_f} \quad (6)$$

333 where N_f and N'_f note the total number of features and the final number of features after reduction,
 334 respectively.

335 The RD of Tamura, LBP, GLRLM, auto-correlation, GLCM, Gabor filters, Haar, Db3 and Db4
 336 are: 50%, 57%, 46%, 50%, 58%, 53%, 42%, 47% and 43%, respectively when using the GA, and 56%,
 337 50%, 49%, 50%, 50%, 48%, 50%, 45% and 52%, respectively when using the RA on a document of the
 338 testing dataset. As a consequence, we conclude that using a feature selection algorithm helps to reduce
 339 the dimensionality of the data, which entails lower computational cost in terms of lighter memory
 340 consumption, processing time and numerical complexity.

341 It is inherently a subjective evaluation to use a visual inspection of the pixel-labeling results of a
 342 texture-based method to draw some conclusions about which set of texture features deduced by using
 343 a feature selection algorithm is well suited for historical DIA. Thus, in this study several per-pixel
 344 and per-block accuracy metrics, namely, the silhouette width (SW) [16], purity per-block (PPB) [11],
 345 and F-measure (F) [15], have been computed based on the defined pixel-accurate ground truth of the
 346 analyzed images of the *HBR2013 dataset*.

347 The silhouette width (SW) assesses the pixel-labeling quality by computing the level of data
 348 compactness and separation based on the intrinsic information concerning the distribution of the
 349 observations into different clusters. The purity per-block (PPB) measures the homogeneity rate of
 350 regions by evaluating the matching regions between the defined pixel-based ground truth and the
 351 obtained pixel-labeling results. The F-measure (F) assesses both the homogeneity and the completeness
 352 criteria of the pixel-labeling results by computing a score resulting from the combination of the precision
 353 and recall accuracies. SW , PPB , and F are computed. The higher the values of the computed metrics,
 354 the better the results. In Table 2, we have used three different colors (red, green, and blue), to quote the
 355 highest SW , PPB , and F values deduced by comparing the performances of each accuracy measure for
 356 each texture-based feature set in the following three cases: with full texture feature set, with texture
 357 features selected using the GA, and with texture features selected using the RA.

358 Good performance has been noted for documents of the training dataset when analyzing the
 359 selected texture features by means of the GA and particularly the Gabor features. However, there is no
 360 significant improvement in performance for documents of the testing dataset due to the complexity
 361 and the wide variety of layouts of the *HBR2013 dataset*. This confirms our observation about the
 362 need to train on documents having similar characteristics in terms of the layout structure and/or
 363 typographic/graphical properties of the historical document image content.

364 To highlight the similarities of the behavior of the different evaluated texture features according to
 365 the use of a full texture feature set, the use of a subset of texture features selected by means of the GA,
 366 and the use of a subset of texture features selected by means of the RA, the correlation analyses of the
 367 F-measure performance of each texture-based feature set are illustrated in Figures 8[a], 8[b], and 8[c],
 368 respectively. Each figure represents a matrix of plots showing the different Pearson's linear correlations
 369 among pairs of the nine texture-based feature sets (Tamura, LBP, GLRLM, auto-correlation, GLCM,
 370 Gabor, Haar, Db3 and Db4). Histograms of the nine evaluated texture-based feature sets appear along
 371 the matrix diagonal, while scatter plots of the texture-based feature set pairs appear in the off-diagonal.
 372 Each dot in each correlation plot represents one historical document image of the testing dataset of the
 373 *HBR2013 dataset*. The displayed Pearson's linear correlation coefficients in the scatter plots highlighted
 374 indicate which pairs of texture-based feature sets have correlations significantly different from zero
 375 (equal to the slopes of the least-squares reference lines in red).

376 Table 3 summarizes the minimum, average, and maximum Pearson's linear correlation coefficient
 377 values of the F-measure performance of pairs of texture-based feature sets according to the use of a full

378 texture feature set, the use of a subset of texture features selected by means of the GA, and the use of a
379 subset of texture features selected by means of the RA.

380 By comparing the different correlation plots and obtained Pearson's linear correlation coefficients
381 when using the full texture feature set, the subset of texture features selected by means of the GA, and
382 the subset of texture features selected by means of the RA, we observe that the Gabor and the three
383 wavelet-based approaches are still highly correlated even if a feature selection algorithm is introduced.
384 This confirms that by using a feature selection algorithm in the Gabor and wavelet approaches only a
385 small subset of relevant features from the original large set of features characterizing the localization
386 of the spatial frequency of a texture have been selected. Nevertheless, we observe higher correlation
387 coefficient values between the Tamura and other investigated features on the one hand and between
388 the LBP and other investigated features on the other hand when selecting features by means of the
389 GA and the RA. This confirms that by using a feature selection algorithm a significant number of
390 texture features which are redundant or irrelevant have been removed. An interesting conclusion
391 that can be deduced from the correlation plots in Figure 8, is that combining the different selected
392 texture feature sets can significantly improve the pixel-labeling quality. Indeed, each feature set has its
393 own particularities. For instance, since Gabor filters is known to be sensitive to the stroke width, they
394 have the advantage to present the best performance in discriminating text in a variety of situations
395 of different fonts and scales. On the other side, the auto-correlation feature set has the advantage
396 to present the best performance for segmenting the graphical contents from textual ones since it
397 highlights interesting information concerning the principal orientations and periodicities of texture
398 [12]. Therefore, combining the different selected texture features from the auto-correlation and Gabor
399 descriptors can be more adequate for segmenting the graphical contents from textual ones on the one
400 hand, and discriminating text in a variety of situations of different fonts and scales on the other hand.

Table 2. Dimensionality and performance evaluation of each texture-based feature set for documents of the training and testing datasets in the following three cases: with full texture feature set, with texture features selected using the genetic algorithm (GA), and with texture features selected using the ReliefF algorithm (RA), using the proposed pixel-labeling scheme on the *HBR2013 dataset*. Internal and external accuracy measures are computed, silhouette width (SW), purity per-block (PPB) and F-measure (F). N_f and N'_f note the total number of features and the final number of features after reduction, respectively. The higher the values of the internal and external accuracy measures, the better the pixel-labeling performances. For each table (*i.e.* the training and testing datasets), the values which are quoted in **red**, **green**, and **blue** colors, are considered as the highest **SW**, **PPB**, and **F** values, respectively by comparing the performances of each accuracy measure for each texture-based feature set in the following three cases: with full texture feature set, with texture features selected using the GA, and with texture features selected using the RA.

		<i>Training dataset</i>								
		Tamura	LBP	GLRLM	Auto-correlation	GLCM	Gabor	Haar	Db3	Db4
Full texture feature set	SW	0.35	0.21	0.13	0.07	0.21	0.26	0.26	0.31	0.28
	PPB	0.71	0.78	0.79	0.73	0.84	0.90	0.80	0.79	0.81
	F	0.38	0.38	0.35	0.43	0.42	0.52	0.45	0.46	0.46
	N_f	16	40	176	20	72	192	80	80	80
Texture features selected using the GA	SW	0.40	0.04	0.14	0.12	0.20	0.29	0.29	0.28	0.31
	PPB	0.73	0.72	0.82	0.80	0.87	0.92	0.83	0.83	0.84
	F	0.39	0.36	0.35	0.42	0.42	0.54	0.45	0.45	0.46
	N'_f	8	20	88	10	36	96	40	40	40
Texture features selected using the RA	SW	0.30	0.07	0.14	0.11	0.26	0.26	0.24	0.29	0.28
	PPB	0.73	0.74	0.77	0.77	0.85	0.88	0.81	0.83	0.83
	F	0.40	0.38	0.34	0.42	0.42	0.49	0.43	0.43	0.44
	N'_f	8	20	88	10	36	96	40	40	40

		<i>Testing dataset</i>								
		Tamura	LBP	GLRLM	Auto-correlation	GLCM	Gabor	Haar	Db3	Db4
Full texture feature set	SW	0.38	0.33	0.35	0.17	0.30	0.28	0.30	0.34	0.30
	PPB	0.77	0.83	0.82	0.81	0.86	0.91	0.83	0.83	0.84
	F	0.40	0.39	0.37	0.43	0.43	0.52	0.44	0.45	0.46
	N_f	16	40	176	20	72	192	80	80	80
Texture features selected using the GA	SW	0.42	0.01	0.42	0.24	0.35	0.33	0.29	0.33	0.36
	PPB	0.76	0.72	0.85	0.85	0.85	0.91	0.86	0.85	0.85
	F	0.40	0.37	0.37	0.43	0.41	0.51	0.43	0.42	0.43
	N'_f	8	17	95	10	30	90	46	42	45
Texture features selected using the RA	SW	0.35	0.15	0.39	0.22	0.36	0.28	0.31	0.34	0.37
	PPB	0.79	0.76	0.81	0.80	0.87	0.89	0.85	0.85	0.86
	F	0.41	0.38	0.36	0.42	0.43	0.49	0.42	0.42	0.42
	N'_f	7	20	89	10	36	98	40	44	38

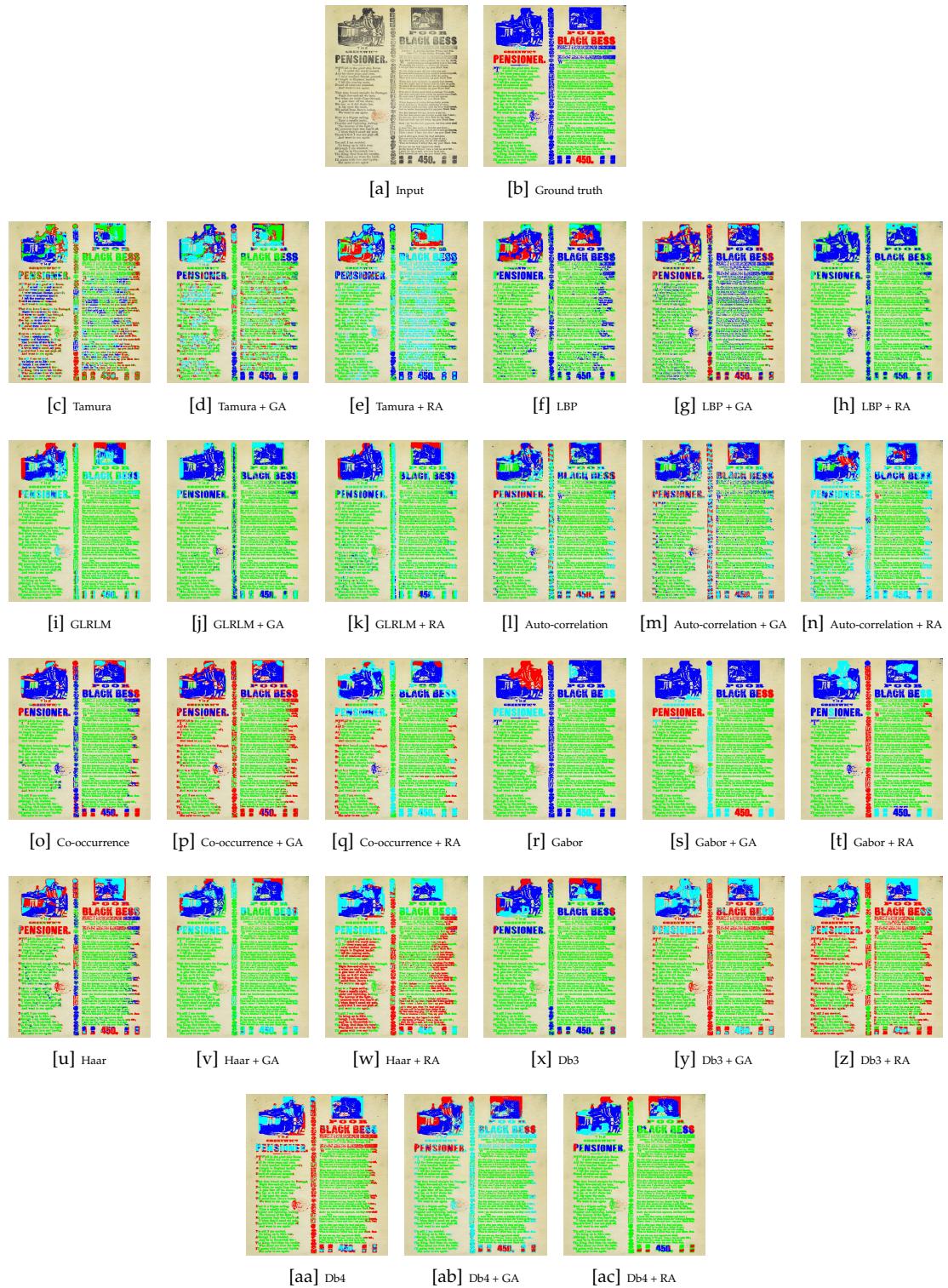
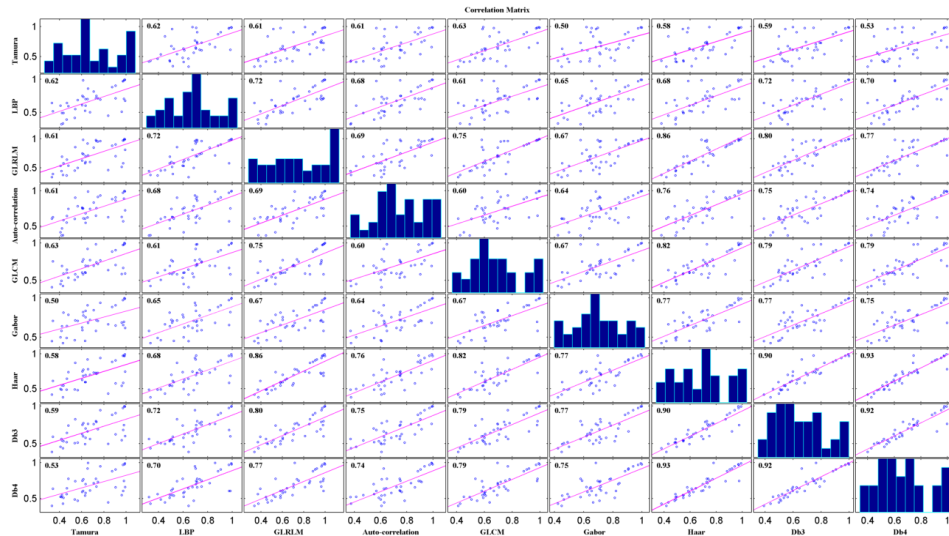


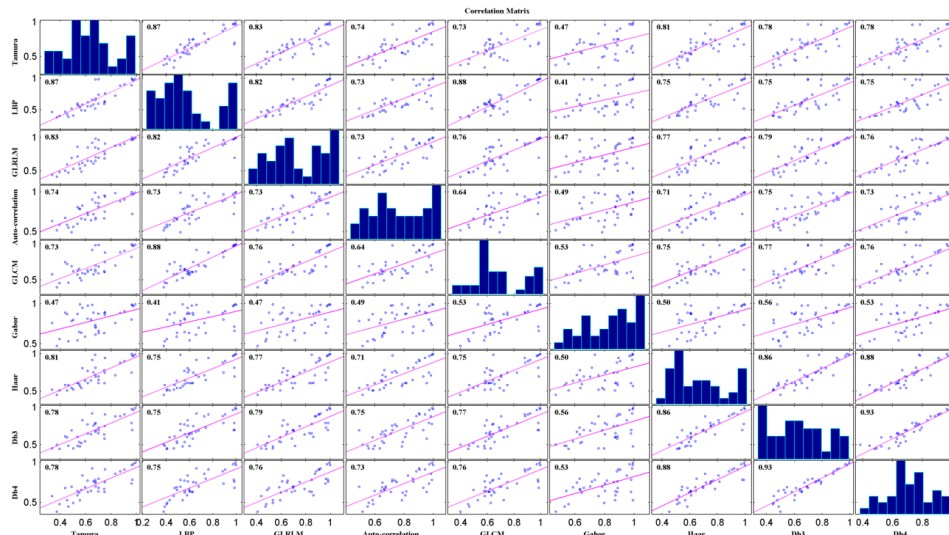
Figure 6. Resulting pixel-labeling images without and with using a feature selection algorithm on a historical document image of the “Three fonts and graphics” category from the training dataset of the *HBR2013 dataset*. The number of class labels is equal to 4. Since the pixel-labeling task is unsupervised, the colors attributed to text or graphics may differ from one document to another. Figures 6[a] and 6[b] illustrate the input image and its associated ground truth, respectively. The remaining figures depict the resulting pixel-labeling images.



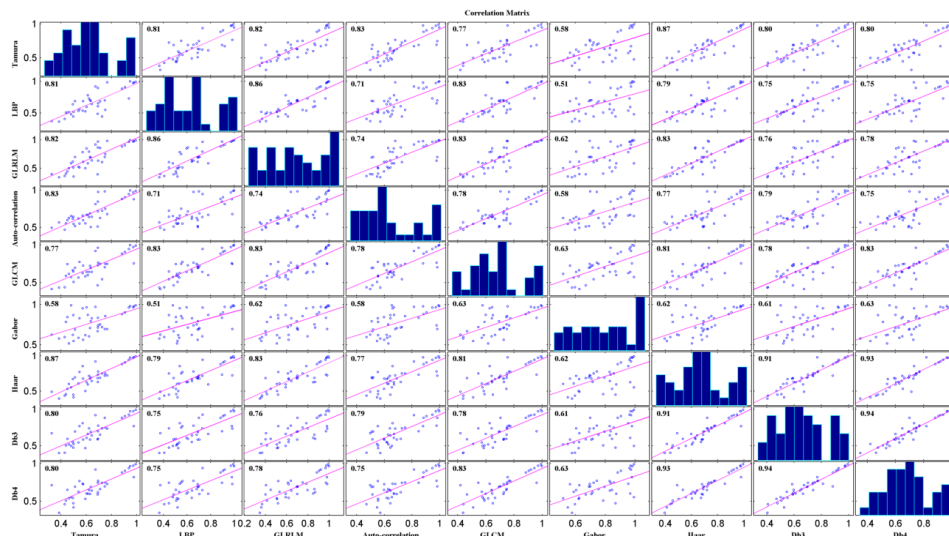
Figure 7. Resulting pixel-labeling images without and with using a feature selection algorithm on a historical document image of the “Three fonts and graphics” category from the testing dataset of the *HBR2013 dataset*. The number of class labels is equal to 4. Since the pixel-labeling task is unsupervised, the colors attributed to text or graphics may differ from one document to another. Figures 7[a] and 7[b] illustrate the input image and its associated ground truth, respectively. The remaining figures depict the resulting pixel-labeling images.



[a] Full texture feature set



[b] Texture features selected using the GA



[c] Texture features selected using the RA

Figure 8. Correlation analysis of the F-measure performance of each texture-based feature set according to the use of a full texture feature set (cf. Figure 8[a]), the use of a subset of texture features selected by means of the GA (cf. Figure 8[b]), and the use of a subset of texture features selected by means of the RA (cf. Figure 8[c]).

Table 3. Minimum, average, and maximum Pearson’s linear correlation coefficient values of the F-measure performance of pairs of texture-based feature sets according to the use of a full texture feature set, the use of a subset of texture features selected by means of the GA, and the use of a subset of texture features selected by means of the RA.

Features	Full texture feature set			Texture features selected using the GA			Texture features selected using the RA		
	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum
Tamura	0.50 (Gabor)	0.58	0.63 (GLCM)	0.47 (Gabor)	0.75	0.87 (LBP)	0.58 (Gabor)	0.78	0.87 (Haar)
LBP	0.61 (GLCM)	0.67	0.72 (GLRLM)	0.41 (Gabor)	0.74	0.88 (GLCM)	0.51 (Gabor)	0.75	0.86 (GLRLM)
GLRLM	0.61 (Tamura)	0.73	0.86 (Haar)	0.47 (Gabor)	0.74	0.83 (Tamura)	0.62 (Gabor)	0.78	0.86 (LBP)
Auto-correlation	0.60 (GLCM)	0.68	0.76 (Haar)	0.49 (Gabor)	0.69	0.75 (Db3)	0.58 (Gabor)	0.74	0.83 (Tamura)
GLCM	0.60 (Auto)	0.7	0.82 (Haar)	0.53 (Gabor)	0.72	0.88 (LBP)	0.63 (Gabor)	0.78	0.83 (LBP)
Gabor	0.50 (Tamura)	0.67	0.77 (Db3)	0.41 (LBP)	0.49	0.56 (Db3)	0.51 (LBP)	0.59	0.63 (GLCM)
Haar	0.58 (Tamura)	0.78	0.93 (Db4)	0.50 (Gabor)	0.75	0.88 (Db4)	0.62 (Gabor)	0.81	0.93 (Db4)
Db3	0.59 (Tamura)	0.78	0.92 (Db4)	0.56 (Gabor)	0.77	0.93 (Db4)	0.61 (Gabor)	0.79	0.94 (Db4)
Db4	0.53 (Tamura)	0.76	0.93 (Haar)	0.53 (Gabor)	0.76	0.93 (Db3)	0.63 (Gabor)	0.8	0.94 (Db3)

5. Conclusions and further work

This paper has presented a comparative study of using two conventional feature selection algorithms for selecting a number of commonly and widely used texture features. This comparative study has been conducted on the *HBR2013 dataset*, using a classical pixel-labeling scheme based on analyzing and selecting features. The proposed pixel-labeling scheme integrates a feature selection step, which has been applied on a training set of the *HBR2013 dataset* in order to select the most relevant texture features of each analyzed texture-based feature set.

We conclude that the performance of a particular feature selection algorithm is highly dependent upon the used texture features. It is admittedly that the proposed pixel-labeling scheme selects fewer texture features with comparable performance. This study has shown that when the numerical complexity and pixel-labeling quality are taken into account, good performance has been noted for documents of the training dataset when analyzing the selected texture features by means of the genetic algorithm and particularly the Gabor features. These results could be explained by the fact that using the genetic operators (such as the crossover and mutation operators) in the GA, guarantee a high diversity of the succeeding populations, and thus more immune to be trapped in a local optima and faster in reaching the global optima. Moreover, the Gabor features perform better than the other features, since they characterize specifically the orientation and spatial frequency of a texture without taking into account the spatial relationships between pixels as concluded in [12].

However, it is not the case for documents of the testing dataset; there is no significant improvement in performance due to the complexity and the wide variety of contents and layouts of the *HBR2013 dataset*. Indeed, it is worth noting that there is awareness that we need a larger database containing documents having similar characteristics in terms of the layout structure and/or typographic/graphical properties of the historical document image content in order to train the different feature selection algorithms. Thus, conducting this study on a larger public annotated dataset of historical books such

425 as the *HBA dataset*⁴ is among the first aspect of our future work. Finally, we intend to extend our
 426 investigation to recent feature selection algorithms.

427 **Acknowledgments:** This study was supported by the LATIS Laboratory – Sousse University and LITIS Laboratory
 428 – Normandie University, which are gratefully acknowledged. The authors would like also to thank Christos
 429 Papadopoulos for providing access to the *HBR2013 dataset*.

430 **Author Contributions:** Maroua Mehri, Ramzi Chaieb, Karim Kalti and Pierre Héroux contributed to the algorithm
 431 design and paper drafting. Maroua Mehri and Ramzi Chaieb designed the experiments and implemented the
 432 algorithms. Rémy Mullot and Najoua Essoukri Ben Amara assembled and proofread the final version of the
 433 manuscript.

434 **Conflicts of Interest:** The authors declare no conflict of interest.

435 Abbreviations

436 The following abbreviations are used in this manuscript:

437	Db3	3-level wavelet transform using 3-tap Daubechies filter
	Db4	3-level wavelet transform using 4-tap Daubechies filter
	DIA	Document image analysis
	F	F-measure
	GA	Genetic algorithm
	GLCM	Gray-level co-occurrence matrix
	GLRLM	Gray-level run-length matrix
438	Haar	3-level Haar wavelet transform
	HAC	Hierarchical ascendant classification
	HBR	Historical book recognition
	LBP	Local binary patterns
	RA	Relief algorithm
	SW	Silhouette width
	PPB	Purity per-block

439 References

- 440 1. A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "Historical document layout analysis
 441 competition," in *International Conference on Document Analysis and Recognition*, 2011, pp. 1516–1520.
- 442 2. A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "ICDAR 2013 Competition on
 443 Historical Book Recognition (HBR 2013)," in *International Conference on Document Analysis and Recognition*,
 444 2013, pp. 1459-1463.
- 445 3. J. Beyerer, F. León, and C. Frese, "Texture analysis," in *Machine Vision*, 2016, pp. 649–683.
- 446 4. J. Calvo-Zaragoza, F. J. Castellanos, G. Vigiensoni, and I. Fujinaga, "Deep neural networks for document
 447 processing of music score images," *Applied Sciences*, 8(5), 654, 2018.
- 448 5. K. Chen, M. Seuret, J. Hennebert, and R. Ingold, "Convolutional neural networks for page segmentation
 449 of historical document images," in *International Conference on Document Analysis and Recognition*, 2017, pp.
 450 965–970.
- 451 6. J. Dubuf, M. Kardan, and M. Spann, "Texture feature performance for image segmentation," *Pattern
 452 Recognition*, pp. 291–309, 1990.
- 453 7. R. Duda, P. Hart, and D. Stork, *Pattern classification*. Second Edition Wiley-Interscience, 2001.
- 454 8. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning
 455 Research*, pp. 1157–1182, 2003.
- 456 9. N. Journet, J. Ramel, R. Mullot, and V. Eglin, "Document image characterization using a multiresolution
 457 analysis of the texture: application to old documents," *International Journal of Document Analysis and
 458 Recognition*, pp. 9–18, 2008.

⁴ <http://icdar2017hba.litislabs.eu/>

- 459 10. K. Kise, "Page segmentation techniques in document analysis," in *Handbook of Document Image Processing and*
460 *Recognition*, 2014, pp. 135–175.
- 461 11. M. Mehri, P. Gomez-Krämer, P. Héroux, A. Boucher, and R. Mullot, "A texture-based pixel labeling approach
462 for historical books," *Pattern Analysis and Applications*, pp. 325–364, 2017.
- 463 12. M. Mehri, P. Héroux, P. Gomez-Krämer, and R. Mullot, "Texture feature benchmarking and evaluation for
464 historical document image analysis," *International Journal of Document Analysis and Recognition*, pp. 1–35,
465 2017.
- 466 13. O. Okun and M. Pietikäinen, "A survey of texture-based methods for document layout analysis," in *Workshop*
467 *on Texture Analysis in Machine Vision*, 1999, pp. 137–148.
- 468 14. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency,
469 maxrelevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp.
470 1226–1238, 2005.
- 471 15. D. Powers, "Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation,"
472 *Journal of Machine Learning Technologies*, pp. 37–63, 2011.
- 473 16. P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of*
474 *Computational and Applied Mathematics*, pp. 53–65, 1987.
- 475 17. Y. Sun, X. Lou, and B. Bao, "A novel Relief feature selection algorithm based on mean-variance model,"
476 *Journal of Information & Computational Science*, pp. 3921–3929, 2011.
- 477 18. M. Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*,
478 pp. 23–69, 2003.
- 479 19. D. Tao, L. Jin, S. Zhang, Z. Yang, and Y. Wang, "Sparse discriminative information preservation for Chinese
480 character font categorization," *Neurocomputing*, pp. 159–167, 2014.
- 481 20. F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block segmentation and text extraction in mixed text/image
482 documents," *Computer Graphics and Image Processing*, pp. 375–390, 1982.
- 483 21. J. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical*
484 *Association*, pp. 236–244, 1963.
- 485 22. H. Wei, K. Chen, A. Nicolaou, M. Liwicki, and R. Ingold, "Investigation of feature selection for historical
486 document layout analysis," in *International Conference on Image Processing Theory, Tools and Applications*, 2014,
487 pp. 1–6.
- 488 23. H. Wei, M. Seuret, K. Chen, A. Fischer, M. Liwicki, and R. Ingold, "Selecting autoencoder features for layout
489 analysis of historical documents," in *International Workshop on Historical Document Imaging and Processing*,
490 2015, pp. 55–62.
- 491 24. H. Wei, M. Seuret, M. Liwicki, R. Ingold, and P. Fu, "Selecting fine-tuned features for layout analysis of
492 historical documents," in *International Conference on Document Analysis and Recognition*, 2017, pp. 281–286.
- 493 25. B. Xue, M. Zhang, W. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature
494 selection," *IEEE Transactions on Evolutionary Computation*, pp. 606–626, 2016.
- 495 26. D. Zongker and A. Jain, "Algorithms for feature selection: An evaluation," in *International Conference on*
496 *Pattern Recognition*, 1996, pp. 18–22.



497 © 2018 by the authors. Submitted to *J. Imaging* for possible open access publication
under the terms and conditions of the Creative Commons Attribution (CC BY) license
498 (<http://creativecommons.org/licenses/by/4.0/>).