A Robust Two Level Classification Algorithm for Text Localization in Documents

R. Kandan, Nirup Kumar Reddy, K.R. Arvind, and A.G. Ramakrishnan

MILE Laboratory, Electrical Engineering Department, Indian Institute of Science, Bangalore 560 012, INDIA
kandan.r@gmail.com, nirupreddy@gmail.com, arvindkr@cmu.edu, ramkiag@ee.iisc.ernet.in

Abstract. This paper describes a two level classification algorithm to discriminate the handwritten elements from the printed text in a printed document. The proposed technique is independent of size, slant, orientation, translation and other variations in handwritten text. At the first level of classification, we use two classifiers and present a comparison between the nearest neighbour classifier and Support Vector Machines(SVM) classifier to localize the handwritten text. The features that are extracted from the document are seven invariant central moments and based on these features, we classify the text as hand-written. At the second level, we use Delaunay triangulation to reclassify the misclassified elements. When Delaunay triangulation is imposed on the centroid points of the connected components, we extract features based on the triangles and reclassify the misclassified elements. We remove the noise components in the document as part of the pre-processing step.

1 Introduction

Most document images invariably consist of a mixture of machine printed elements such as logos, text, barcodes etc and handwritten elements such as address or name texts, signatures, markings etc. From established methods of machine-printed and handwritten character recognition it is understood that the methods are quite different from each other. Hence a necessary preprocessing step to an OCR is the separation of machine-printed and hand-written elements.

Imade et al. [1] have described a method to segment a Japanese document into machine-printed Kanji and Kana, handwritten Kanji and Kana, photograph and printed image. They extracted the gradient and luminance histogram of the document image and used a feed forward neural network in their system. Kuhnke et al. [2] developed a method for the distinction between machine-printed and handwritten character images using directional and symmetrical features as the input of a neural network. Guo and Ma [3] have propose a scheme which combined the statistical variations in projection profiles with hidden Markov models (HMMs) to separate the handwritten material from the machine printed

G. Bebis et al. (Eds.): ISVC 2007, Part II, LNCS 4842, pp. 96–105, 2007.

text. Fan et al. [4] have proposed a scheme for classification of machine-printed and handwritten texts. They used spatial features and character block layout variance as the prime features in their approach. They have also claimed that this technique could be applied to English or Chinese document images. Pal and Chaudhuri [5] have used horizontal projection profiles for separating the printed and hand-written lines in Bangla script. In this paper we use a set of seven 2D invariant moments, that are insensitive to translation, scale, mirroring and rotation as the feature for distinguishing the printed and handwritten elements. Then we go on to use Delaunay triangulation to reassign the labels assigned to the elements. We find that the accuracy achieved is around 87.85% using the nearest neighbour classifier and with the SVM classifier the accuracy is 93.22%

2 System Description

The entire system is divided into three stages. The first stage is the preprocessing stage, in which the document is cleaned of all the noise components present such as spurious dots and lines. In the second stage we extract the features based on invariant moments for classification of the elements into printed or handwritten. This classification is done with the Nearest neighbor and SVM classifiers separately. Finally, in the third stage we use Delaunay Triangulation to reassign the class labels assigned to the elements. The flowchart of the entire system is as shown in Figure 1. Let us look at the three stages in detail.

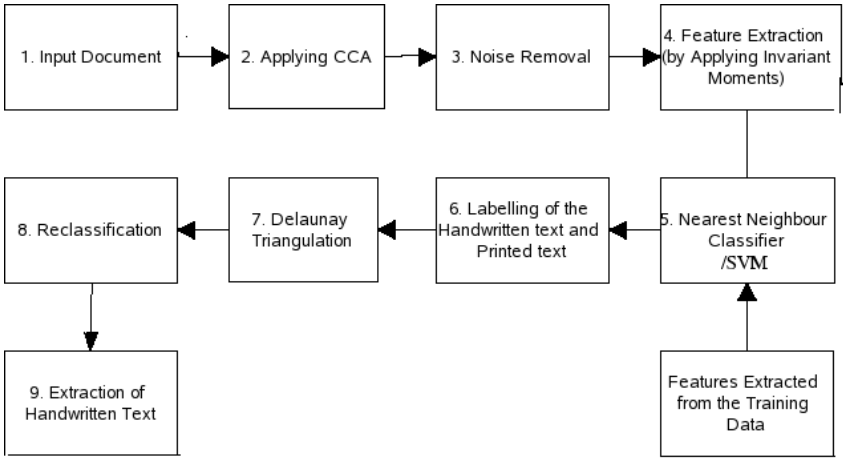


Fig. 1. Flowchart showing the Algorithm for Text Localization

2.1 Preprocessing

In this stage we remove the noise elements such as dots and lines in the document image. This process is described below.

- Apply Connected Component Analysis (CCA)
- Obtain Bounding Boxes of the connected components. Then we find the area of each connected element A_i and also find the minimum area A_{min} , maximum area A_{max} in the entire document.
- If one of the following conditions is met then it indicates noise in the document and it is removed.
  - If the value of; $(A_i\text{- } A_{min})/(A_{max}\text{- } A_{min}) < T_1$ (a threshold value, for the test documents it is set at 0.002)
  - Aspect ratio is used to remove horizontal and vertical lines.
  - Also if the height or width of the connected component is lesser than a threshold value, then it is a spurious element or a dot, and we remove it.

The thresholds which are used have been chosen empirically.

2.2 Feature Extraction

We use the features drawn by invariants moment technique which is used to evaluate seven distributed parameters of an image element. The invariant moments (IMs) are well known to be invariant under translation, scaling, rotation and reflection [6,7]. They are measures of the pixel distribution around the centre of gravity of the character and allow to capture the global character shape information. In the present work, the moment invariants are evaluated using central moments of the image function f(x,y) up to third order.

Discrete representation of the central moments is;

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \tag{1}$$

where for p,q = 0,1,2,... and x and y are moments evaluated from the geometrical moments of M_{pq} as follows

$$\bar{x} = \frac{M_{10}}{/} M_{00} \quad and \quad \bar{y} = \frac{M_{01}}{M_{00}} \tag{2}$$

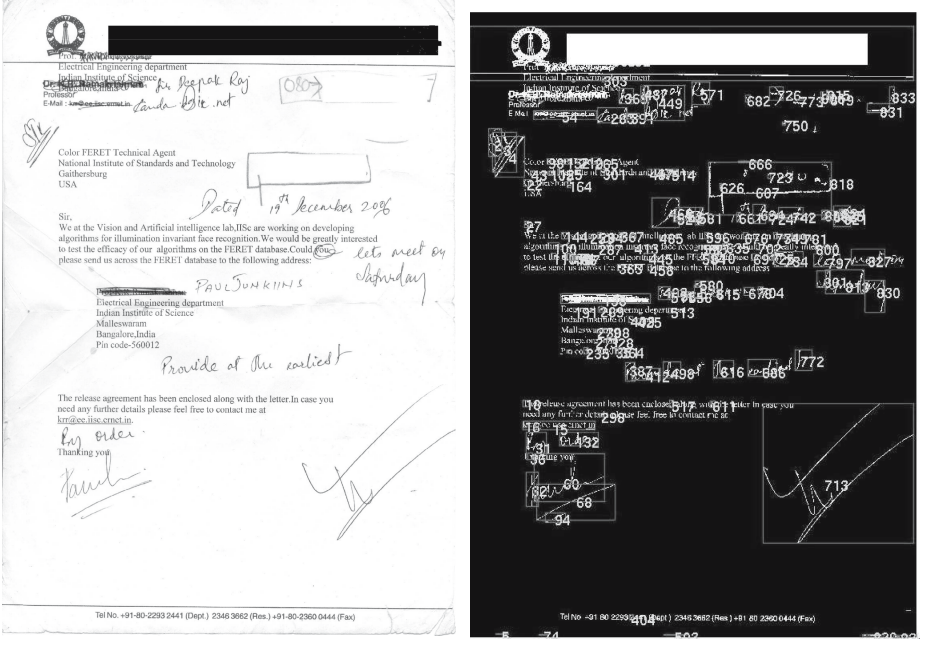
$$M_{pq} = \sum_x \sum_y (x)^p (y)^q f(x, y) \tag{3}$$

The central mmoments $\mu_{00}, \mu_{10}, \mu_{01}, \mu_{11}, \mu_{20}, \mu_{02}, \mu_{30}, \mu_{03}, \mu_{21}, \mu_{12}$ of order upto 3 are calculated. A further normalization for variations in scale is implemented using the formula'

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}} \tag{4}$$

From the central moments, the following values are calculated;

$$\phi_1 = \eta_{20} + \eta_{02}$$
$$\phi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{02}^2$$
$$\phi_3 = (\eta_{30} + 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$
$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} - \eta_{03})^2$$
$$\phi_5 = (\eta_{30} + 3\eta_{12})^2(\eta_{30} - \eta_{12})(\eta_{30} - 3\eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2$$
$$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[(3(\eta_{30} - \eta_{12})^2 - (\eta_{21} - \eta_{03})^2$$
$$\phi_6 = (\eta_{20} - \eta_{02})(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 +$$
$$\eta_{11}(\eta_{30} - \eta_{12})$$
$$\phi_7 = (3\eta_{21} - \eta_{02})(\eta_{30} + \eta_{12})(\eta_{30} - \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2 +$$
$$(3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[(3(\eta_{30} - \eta_{12})^2 - (\eta_{21} - \eta_{03})^2$$



(a) Input Document

(b) Image after Handwritten elements are classified using NN classifier

Fig. 2.

where ϕ_7 is a skew invariant to distinguish mirror images. In the above, ϕ_1 and ϕ_2 are second order moments and ϕ_3 through ϕ_7 are third order moments. ϕ_1 (the sum of the second order moments) may be thought of as the spread of the pattern; whereas ϕ_2 may be interpreted as the "slenderness" of the pattern. The third order moments, ϕ_3 through ϕ_7 do not have any direct physical meaning but includes the spatial frequencies and ranges of the image element.

2.3 Classification

The following classifiers are used to localisze the handwritten text regions; i) Nearest Neighbour Classifier ii) Support Vector Machines.

The feature vector is considered to be a point in the feature space and the training data is a distribution of points in the feature space. Now for the test block, we extract the feature vectors and then the Eculidean distances are calculated from each of the points of the training data. Using the Nearest neighbour classifier principle we assign the class label of the training vector which has the minimum distance from the test vector, to the test block. Figure 2(a) and 2(b) depict the input image from which the handwritten elements are to be separated and the output after classification using the nearest neighbour classifier. In Figure 2(b), all those elements that have a class value of two, which represent handwritten text are marked by a magenta Bounding Box.

We trained an SVM classifier with Radial Basis Function(RBF) kernel[8] with the invariant moments as the features. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameter of the error term. The RBF kernel used is given below

$$K(x_i; x_j) = exp(-\gamma(||x_i - x_j||)^2, \gamma > 0 \tag{5}$$

The kernel parameters were chosen by using a five fold cross validation using all the training data samples. The best parameters were found to be C= 256 and γ = 0.0625 which was operating at an efficiency of 96%. With a lower computation cost i.e, by reducing the value of C=32 we achieve an efficiency of 95% when cross validation is done as shown in Figure 3.

Figure 4 depict the image in which the handwritten elements are separated using SVM's.

2.4 Reclassification

We now use Delaunay triangulation[9] for reclassification of the elements which are misclassified. We briefly give the definition of Delaunay triangulation. Delaunay triangulation of a set of non-degenerate vertices V is defined as the unique triangulation with empty circles, i.e, no vertex lies inside the circumscribing circle of any Delaunay triangles, as follows:

$$DT(V) = (p_i.p_j, p_k)\epsilon V^3, B(p_i.p_j, p_k) \cap V \setminus (p_i.p_j, p_k)\epsilon\phi \tag{6}$$

where B($p_i; p_j; p_k$) is the circle circumscribed by the three vertices $p_i; p_j; p_k$ that form a Delaunay triangle. The Delaunay triangulation carried out on printed text/handwritten text regions have the following features:

– The lengths of the sides of most triangles in a printed text region are similar as compared to the lengths of the handwritten text.

- Triangles in the printed text have their longest and similar sides link the point pairs between two adjacent text lines above which is also printed text.
- The height of the triangles in the printed text region are uniform.

The above features are extracted after applying the Delaunay triangulation and a threshold value based on comparing with the neighbouring points is set to reclassify the text as machine printed or handwritten text. If a particular element and its neighbouring elements have similar features and the element in consideration is labelled differently then it is assigned the label of the neighbours.

We carry out the reclassification as per the following algorithm:

- First the Delaunay triangulation is done on the document for the centroid points as shown in Figure 5(a) in which NN classifier is used and 5(b), in which SVM is used as the classifier.
- Now, let us consider a centroid point P(x,y). A number of triangles originate from P. Thus P is associated with other centroid points P_1, P_2 and P_n by the Delaunay triangles. All such points P_1, P_2 and P_n which are connected to the point P are said to be the neighboring points of P.
- After we get these neighboring points we compare the label of each point with the label of the centroid point P(x,y). If the label is not the same then we find the degree of similarity of the triangles; i.e in this case the difference between the heights of the element defined by the neighboring centroid points and the height of the element defined by the reference point P(x,y). If the difference is less than 7 pixels then we increment a count indicating the similarity in height.
- If more than 50% of the neighbouring points have similar height i.e, the diference in height is less than 7 pixels then we reassign the label i.e 1 as 2 or 2 as 1. Let us assume count as C_i and the total no of neighbouring points as N_i for the centroid point i then if the value of $\frac{C_i}{N_i} * 100 > 50$ - the label is reassigned else the label remains the same. This is then done for all the centroid points.
- This means that if the centroid point P, has different text as compared to the neighboring points, then the height difference will be greater than 7 pixels. However if the centroid point P, has similar text as compared to the neighboring points and is misclassified we compare the height feature of the triangles and compute the difference in height. If the feature is same(i.e if the difference in height is less than 7 pixels), with more than 50% of the neighboring points then the point P(x,y) is given the label of the neighboring points and is re-classified. If the point P has the same label as that of its neighbors then the above steps are not required and the above algorithm is done for the next point.

Figure 6(a) and 6(b) shows the image of the document with the handwritten elements localized within a magenta colored bounding box after the labels are reclassified.

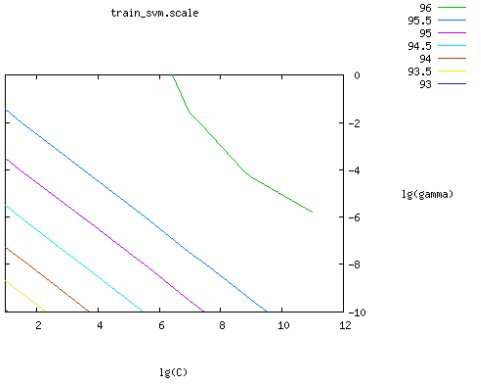


Fig. 3. Graph showing the cross-validation results where γ is varied between 1 to 2_{-10}, C is varied between 1 to 2_{+10} and the best pairs of (γ, C) are choosen

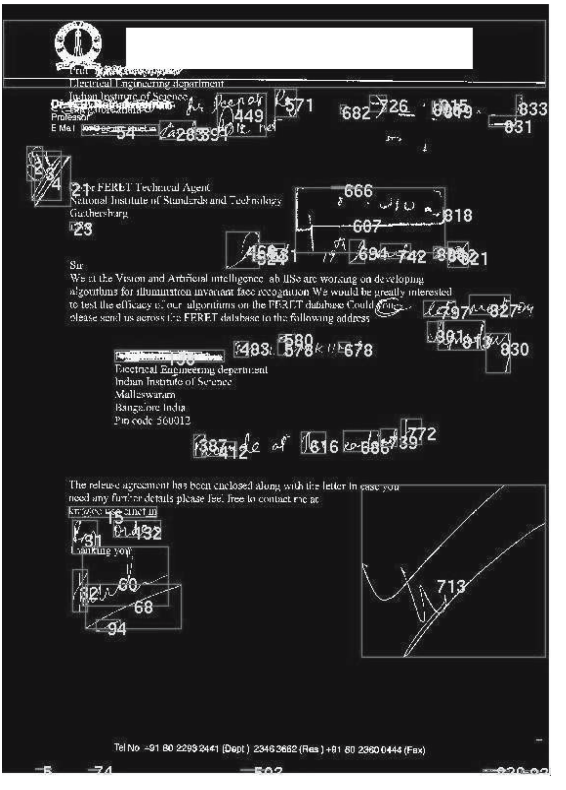
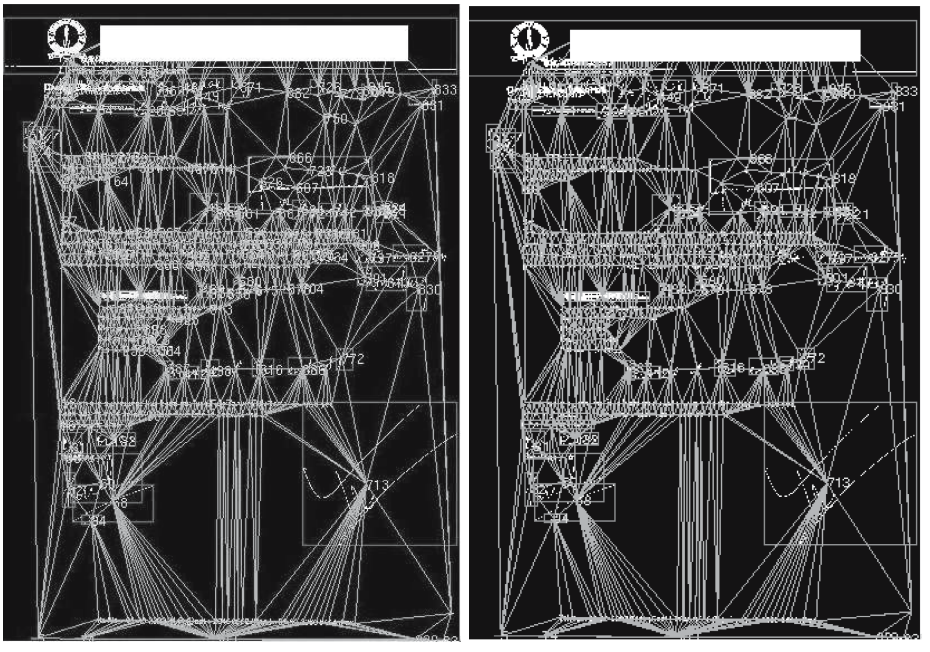


Fig. 4. Image after the Handwritten elements are classified using SVM



(a) Image after the NN classification showing Delaunay triangulation's plot

(b) Image after the SVM classification showing Delaunay trangulation's plot

Fig. 5.

3 Experimental Results

3.1 Data Description

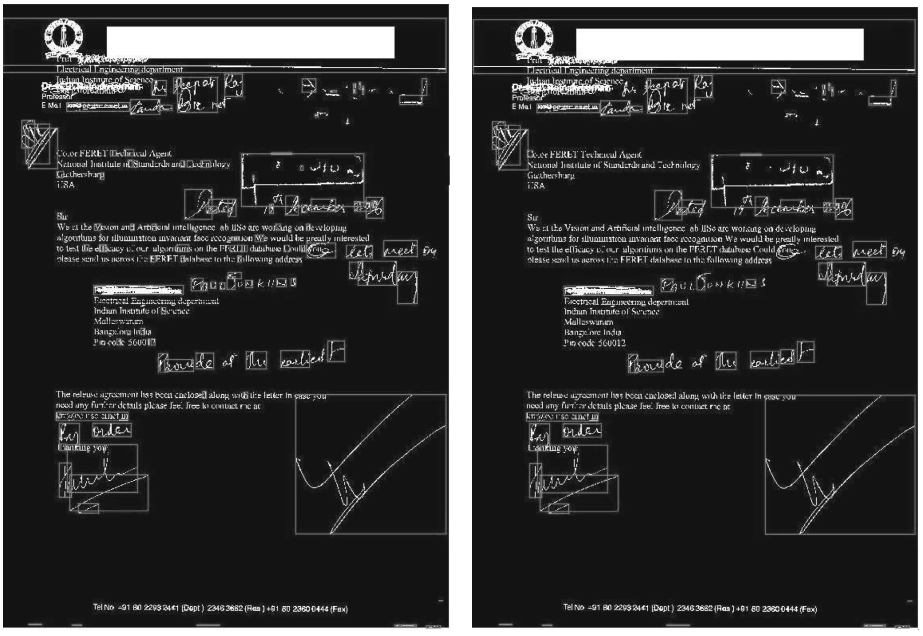
The training data are extracted from over 500 documents which contain predominantly machine printed elements. The handwritten elements are composed of text that is both cursive and block handwriting besides signatures, dates and address locations. Our test data consists of 150 English document images, scanned at 200 dpi and stored in 1-bit depth monochrome format. These documents contain handwritten elements, signatures, logos and other such things along with free-flowing text paragraphs.

3.2 Accuracy Calculation

Table (1) shows the classification accuracy using the proposed method. Using the nearest neighbour classifier we find that the number of printed text elements which are misclassified is higher which reduces the overall accuracy after the second stage of classification. The SVM classifier shows better accuracy and also the misclassified elements are few in number. There were a total of 1,678 handwritten elements in the test documents and the nearest neighbour classified 1,475 elements correctly while the SVM classified 1,565 elements correctly.

Table 1. Classification accuracy of Handwritten text from test data of 150 documents

Localization of Handwritten text	Correctly Classified Elements	Misclassified elements	%Accuracy
using Nearest Neighbour at the first stage of classification	1,475	672	87.85%
using SVM at the first stage of classification	1,565	156	93.22%



(a) Handwritten elements re-classified after the NN classifier in the bounding box

(b) Handwritten elements re-classified after the SVM classifier shown in a bounding box

Fig. 6.

4 Conclusion

In this paper we have presented a novel method for extraction of handwritten text from the machine printed text in the documents. After the first level of classification it is found that there are a lot of printed text which is misclassifed and hence we employ delaunay triangulation to reclassify the text. This is found to give us higher efficiency and accurate results. This classification which is done over two levels improves the overall accuracy of localization of handwritten text rather than a single level of feature extraction and classification. This method

can be used to extract handwritten components from noisy documents and hence it is a robust algorithm. The misclassification of printed text has been greatly reduced by use of SVM as opposed to nearest neighbour classifier. However this method fails to accurately localize when the handwritten elements have similar structure to machine printed text for eg. block letters. Also in some cases when there is a continuous set of handwritten text, only certain part of the handwritten text are separated. This can be addressed by optimal selection of more features at the first level of classification which can be further passed to Delaunay triangulation.

References

1. Imade, S., Tatsuta, S., Wada, T.: Segmentation and Classification for Mixed Text/Image Documents Using Neural Network. In: Proceedings of 2nd ICDAR, pp. 930–934 (1993)
2. Kuhnke, K., Simonicini, L., Kovacs-V, Z.: A System for Machine-Written and Hand-Written Character Distinction. In: Proceedings of 3rd ICDAR, pp. 811–814 (1995)
3. Guo, J.K., Ma, M.Y.: Separating handwritten material from machine printed text using hidden markov models. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 439–443 (2001)
4. Fan, K., Wang, L., Tu, Y.: Classification of Machine-Printed and Hand-Written Texts Using Character Block Layout Variance. Pattern Recognition 31(9), 1275–1284 (1998)
5. Pal, U., Chaudhuri, B.: Automatic Separation of Machine-Printed and Hand-Written Text Lines. In: Proceedings of 5th ICDAR, Bangalore, India, pp. 645–648 (1999)
6. Ramteke, R.J., Mehrotra, S.C.: Feature Extraction Based on Invariants Moment for Handwriting Recognition. In: Proc. of 2nd IEEE Int. Conf. on Cybernetics Intelligent System (CIS2006), Bangkok (June 2006)
7. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, Pearson Education (2002)
8. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines, Software (2001), available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
9. Davoine, F., et al.: Fractal image compression based on Delaunay triangulation and vector quantization. IEEE Trans. Image Process 5(2), 338–346 (1996)