1-1-2015

Template Generation from Postmarks Using Cascaded Unsupervised Learning

Elisa H. Barney Smith
Boise State University

Gernot Fink
TU Dortmund

 doi: 10.1145/2809544.2809555

Template generation from postmarks using cascaded unsupervised learning

Elisa H. Barney Smith
Electrical and Computer Engineering Department
Boise State University
Boise, ID 83725-2075, USA
ebarneysmith@boisestate.edu

Gernot Fink
Department of Computer Science
TU Dortmund
Dortmund, Germany
gernot.fink@tu-dortmund.de

ABSTRACT

Information in historical datasets comes in many forms. We are working with a set of World War I era postcards that contain hand written text, some preprinted text, postage stamps and postmark/cancellation stamps. The postmarks are of considerable interest to collectors looking for images of samples they had not previously seen. The postmarks also provide information on the originating location of the card that complements the information in the address block.

The postmarks vary considerably with towns and dates, but also styles. The styles can be grouped into categories. A method for automatically extracting templates for each category of these postmark stamps is described. The problem is complicated by the high levels of degradation present in the cards. The approach uses a cascade of unsupervised learning steps separated with image cleaning. This introduces averaging steps, which reduces noise. It also provides a reduction in the number of comparisons between samples. While merges happen at each stage, the number of times merges are needed within each stage is reduced. The templates once extracted can be used to group the postmarks, and will contribute information about the postmark content to better separate the postmark from the paper and other interfering marks to extract further information about the postmarks and postcards.

Keywords

Image clustering, sequential learning, document seal recognition

1. INTRODUCTION

The First World War brought a substantial development in the military postal services which were supplied for communication between soldiers and their relatives and friends at home. On the German side, the tremendous amount of mail, approximately 28.7 billion pieces, was delivered from the front to the homeland and vice versa, of which approximately 25% was postcards [**?**, p. 29]. Therefore, this so-called *feldpost* can be considered as the equivalent of today's social media. Analyzing collections of feldpost, consequently, needs to take into account the nature of this high volume data source, even though current collections of WWI feldpost are still limited in their volume (cf. e.g. [**?**]). This means that individual mail pieces will be of limited interest to historians and only the analysis of larger numbers of postcards will be able to reveal insights about socio-historical aspects of the World War I era.

As the automatic transcription of historical documents in general and feldpost postcards in particular is still infeasible today, it is desirable to employ automatic analysis tasks that assist experts in the process of making this valuable information source accessible. Feldpost postcards usually contain some pre-printed material, the handwritten content, usually at least one postmark, and sometimes a postage stamp, see Figure 2a. In this study we focus on the postmarks. These can provide valuable cues for determining the place of origin or the feldpost distribution center from which the card was sent. Some postmarks also identify the military unit to which the sender belonged. A more detailed analysis could also extract the date frequently present in postmarks in order to be able to datemark the respective communication. It is therefore of interest to develop techniques that can separate the handwritten text from the machine printed text on the original postcard, and separate the postage stamp and postmark stamp from the text.

In this paper we propose a method for automatically dividing the postmark samples into groups of like styles and extracting a template for each category. Figure 1(a) shows an example of an ideal (semi-manually generated) exemplar and (b)-(j) show example postmarks that are in the same style and would naturally be grouped together. Samples (b)-(e) contain some common text. (f)-(j) show samples that differ further in content, but are also in the same style. (k)-(o) show samples that are not in the same style, but could be mistakenly included. They all have a pair of horizontal bars crossing the circle, but the semicircle in Subfigure (k) does not continue all the way to the horizontal bars, the horizontal bars in Subfigures (l), (m) and (n) do not continue all the way across the circle. The postmark in Subfigure (o) does not have the semicircle feature. The goal is to automatically create the exemplar in Figure 1(a) and similar exemplars that correspond to the postmark styles in (k)-(o). This is the opposite problem from retrieval where the sample in Figure 1(a) would be used to identify (b)-(j). Here

To appear in the HIP '15: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing

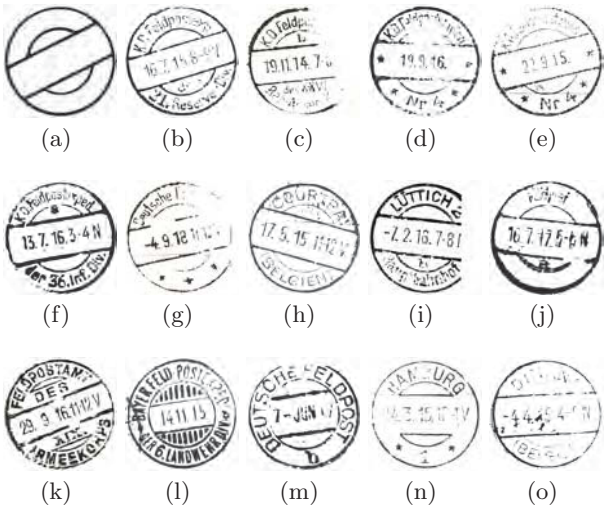


Figure 1: Examples of postmark categorization. (a) the desired template. (b)-(j) samples that belong in that category (k)-(o) samples that are from different categories.

we want to identify that postmarks (b)-(j) should be used to form (a).

The problem is complicated by the high levels of degradation present in the cards, Figures 2 and 4. The ink is faded, and the paper is yellowed. A significant quantity of both occlusion and dropout exists in the same images. The postmarks overlap the other content on the card. The rubber stamp sometimes did not make full contact with the paper or did not physically 100% overlap the postcard paper resulting in an incomplete image. Sometimes there is an excess of ink which connects the text and line components. The postmark ink can be darker, lighter or the same intensity as the handwritten text. The content of the true postmark image, like the date and the military unit or town, can change between samples. The true variations between the postmarks are as frequent as the variations introduced by the noise.

Once the templates are extracted, they can be used to group the postmarks and postcards for analysis. The templates will contribute information about the postmark content, which can be used to extract additional information about the specific postmarks, and to better separate the postmarks from the background.

Section 2 provides an overview of related work and describes related approaches. The details of the technique that was used to prepare the images is described in Section 3. The experimental procedure and results are shown in Section 4. The paper concludes in Section 5.

2. BACKGROUND AND PAST WORK

Segmenting the postcard image is related to segmenting envelope content [**?**, **?**, **?**, **?**]. To increase postal automation, locating and segmenting the destination address block is of greatest interest. Liu [**?**] used OCR to recognize the dates in heavily distorted postmarks. The location of the dates within the image was known aprioi. Compared to the envelopes in these cited studies, the content of the feldpost postcard is much more crowded, and due to its age they are more noisy.

The process of segmenting and recognizing logos found in business letters shares characteristics of the postmark problem [**?**, **?**, **?**]. The templates for each category are usually predefined, but some work aims to find near matches for possible trademark infringement [**?**]. Seal stamps, such as are common in Asia, are more similar to postmarks [**?**].

Logos come in a variety of shapes and sizes. The postmarks in this collection are all round. There are occasionally ($< 0.2\%$) some other stamped marks on the postcards which are rectangle or oval. Those are not being considered in this work.

We relied on the assumption that the postmarks were all circular to locate the postmark in the image. It was automated and the accuracy results are comparable to or better than those reported in the literature for logo or seal segmentation [**?**].

In supervised learning cascading of classifiers is common. Many works use boosting, either with the same classifier architecture, or complementary architectures. Zhu [**?**] used this approach for logo detection.

We desire in the end to produce on the order of 10 style exemplars The method we propose uses a cascade of unsupervised learning steps. It iteratively merges the data to make new approximations of the representative samples at the end of each stage. If clustering were done by a single application of the algorithm, the noise present in the image samples would have a large influence on the result, probably a larger influence than the structural details we desire to capture. Therefore our method is based on an incremental and iterative clustering approach.

Algorithms like C-means iteratively refine the estimate of the cluster center, but in each step use the original image sample. In the proposed algorithm the original postmark images are only available to the first stage. The averaging between the stages is vital to compensate for the high level of structural noise present in the images. This iterative clustering also has the benefits that the quantity of comparisons and merges is decreased. This can be helpful to other domains where the process of calculating the difference between samples or merging the samples is complex. An algorithm such as C-means would have to repeat these difference and merge calculations at each step in the process before the algorithm converges on a final solution.

3. IMAGE PREPARATION

The image preparation process involves an initial enhancement of the image to make the faded text and lines prominent. From this the postmarks are detected with a circle detection algorithm. The enhanced images used for circle detection are combined with Sauvola thresholded images. These are the input to the first level of clustering. Details of these techniques are described next.

3.1 Dataset of Historical Postcards

The dataset of German feldpost postcards considered in this work is part of a private collection of postcards from World War I [**?**]. The collection focuses on mail items related to the western war zone, i.e., postcards sent to and from the German front-lines in Belgium, and northern France. The collection comprises mostly postcards from bequests such that repetitive communication between parties can be ob-

served in the data.

The total collection comprises several thousand postcards that have been acquired at different times and a portion has been digitized. All postcards were photographed at approximately 600 dpi with a total resolution of 4288×2484 pixels. The photographs were taken with a digital SLR camera in front of a red background that is removed by color space thresholding. The postcards themselves are approximately 3200×2100 pixels in size.

In this work, two sub-sets of this collection are considered, which consist of 100 and 460 postcard images, respectively. These datasets were used for evaluation purposes in [**?**]. Many of these cards contain multiple postmarks. After selection constraints were applied, 581 postmark samples were used in this work.

3.2 Initial Enhancement

To better identify the markings on the card, many of which are greatly faded as shown in Figure 2a, a series of pre-processing steps were applied. To compensate for uneven background brightness and stains, the background image was estimated for each color plane [**?**] and was subtracted from the image. Because the postcards are yellowed from age, the blue colorplane contains mostly noise. The images in the red and green color planes were contrast stretched, median filtered and smoothed. These two color planes were then averaged. The faded as well as clean text at this point is well separated from the background, Figure 2b.

3.3 Circle Detection

Circles and arcs were detected [**?**] in a 2x reduced version of the enhanced image. Circles and arcs were restricted to those with radii in the range 250 to 500 pixels, and the total arc length had to exceed 200 pixels. These were heuristically selected thresholds. An example of the circles and arcs that were detected is shown in Figure 2b. The circles and arcs that satisfied the limiting conditions are shown in red. The other circles and arcs that were detected are shown in blue. Detected circles and arcs were checked to see if they were concentric, which can occur for some postmark designs. These were merged to include just the outer circle. A total of 632 potential postmarks resulted. A manual search through the 560 postcard images showed that only 3 postmarks were missed by this method and 5 false positives resulted, Figure 3. This accuracy is comparable to or better than that reported in the literature for circular stamp detection [**?**]. The dataset was not amended to correct for these detection errors.

The postmark images came in a variety of sizes. Several marks have the same basic style, but with different sizes and different text. The sizes of the marks could be used as a feature to distinguish the marks. To get a taxonomy of the different postmark styles, all the postmark images were resized to 380×380, which includes a small buffer around the exterior.

3.4 Further Image Processing

The postmarks were extracted from the original color image. From these the red and green planes were thresholded with Sauvola ($R = 128$, $k = 0.1$, $W = 10$) to provide additional content that might have been smoothed in the earlier preprocessing steps. The two Sauvola images and the thresholded pre-processed image were OR-ed, Figure 4. A 3x3 median filter was applied to remove small specks. Because many postmarks had extremely faded ink, many strokes detected were very thin. Aggressive cleaning would remove much of the content.

(a) (b)

Figure 2: Postcard image (a) original and (b) text enhanced with detected circles and arcs. Red circles and arcs met a validity threshold.

(a) (b) (c) (d) (e)

Figure 3: Examples of false detected postmarks.

3.5 Text Removal

We aim to create a template that will contain arcs and lines indicating the structural components of the postmark, such as in Figure 1a. The text can be an identification feature for grouping postmarks, but it also varies by date and by city or military unit. The characters do not contribute to the template. To emphasize the structural components, a "cleaned" image was created that omitted connected components that were likely text characters. Components with a width and height both less than 45 were designated as possible text. This includes speckle noise. It also removed some small broken strokes. Other content that overlaps the postmark inhibits the removal of text and will remain itself.

Any postmark where the convex hull on the remaining information occupied less than 50% of the postmark area was rejected. These would look mostly like arcs and they would not contribute to forming good templates. After this limiting condition was applied, the set of 632 postmarks was reduced to 581 postmarks. These remaining 581 images were used to form the templates through the iterative clustering process.

4. EXPERIMENTS AND RESULTS

To create a small number of homogeneous clusters, clustering was conducted in a series of stages, instead of one large stage. Use of multiple stages allows us to do intermediate cleaning and noise removal. Particularly the averaging that occurs when symbols are merged at the end of a clustering stage reduces noise and enhances common structural features.

The first stage did a very coarse clustering to find large similarities, but allowed for differences (largely from the dropout and occlusion) that would be "washed away" through averaging when the samples are merged at the end of the

Figure 4: Postmark stamps (a) original, (b) general enhancement, (c) Sauvola from red and green planes, (d) combination of first enhancement and Sauvola, and (e) the "cleaned" postmark images omitting isolated text.

stage. This reduced the size of the dataset, and removed large quantities of noise. The second stage proceeds in several fine steps to gradually refine the clusters, cautiously making merges of similar samples and avoiding false alarms. At each step in the second stage approximately 10% of the clusters were removed.

4.1 Clustering Procedure

The postcards were processed according to the technique described in the Section 3 to extract and prepare postmark samples. The 581 prepared postmarks were input to a cascade of agglomerative clustering algorithms.

Agglomerative hierarchical clustering [**?**] initializes with each symbol in its own cluster. All pairs of clusters, X_i and X_j, are compared to find the pair with the minimum set distance All levels use the set distance metric

$$D_{max}(X_i, X_j) = \max_{x \in X_i, x' \in X_j} ||x - x'||. \quad (1)$$

D_{max} determines the maximum distance between any pair of points between the two sets. Those two clusters, are then combined. D_{max} discourages elongated clusters, making it good for compact equal sized round clusters.

4.2 Matching

Matching was done on the image pixel values. To match the images no translation was needed, but a proper angle of rotation was needed. Each pair of images was tested in the range of rotation angles $-60^\circ \leq \phi \leq 60^\circ$ to find the angle producing the best match.

Figure 5: Output of Level 1 processing.

Depending on the parameters, a few clusters that are large or several clusters that are small will result. Larger clusters will be more diverse, particularly when the input is noisy. We desire to create a small number of homogeneous clusters.

Two image distances were used. In the first level of clustering the distance is the sum of the absolute difference normalized by the number of black pixels in both images:

$$dist_1 = \frac{|img_1 - img_2|}{|img_1| + |img_2|}. \quad (2)$$

In the subsequent levels the likelihood that one of the potential templates was a subset of another was exploited by a distance metric of 1 minus the number of matching pixels between two images normalized by the smaller total number of black pixels:

$$dist_2 = 1 - \frac{|img_1 \bigcap img_2|}{min\{|img_1|, |img_2|\}}. \quad (3)$$

4.3 Results

The first level clustering produced 377 clusters, from the 581 samples. 93 of these contained multiple members. All images that did not merge into a multiple member cluster in the first stage were excluded from contributing to template formation for use in the second stage. They likely contained high levels of noise. The false alarms, Figure 3, were all in this category. It is possible that some rare postmark styles could be in this category and would therefore be falsely rejected. Observations show this is not the case on this dataset. Outliers would produce a template with the noise and text details, and not just the style structure.

The results of the clustering were combined to produce an image representative of the cluster. This should show the structure present in the cluster members and the variation will be eliminated through averaging. The cluster members were all rotated to the angle that makes the image best match the first cluster member. The image intensities were averaged.

Some samples of the output of Level 1 are shown in Figures 5 and 6. While these look very good, several forms appeared multiple times, Figure 6. These samples were thresholded at 80% and results were removed whose convex hull after merging extended over less than 50% of the circle. At this point 86 categories remained. These cluster averages were the input to the second stage.

The next levels of processing are intended to reduce the number of clusters by reducing duplicates and emphasizing common features. They are starting with samples containing more structural content from the averaging of the cluster members. In the second and subsequent clustering stages a threshold to stop clustering and repeat the member averaging was chosen to be at the 10^{th} percentile of the measured distances between all pairs of images. This was repeated

(a) (b) (c) (d) (e)

Figure 6: Output of Level 1 processing showing near duplicates.

(a) (b) (c) (d) (e)

Figure 7: Output of Level 2 processing.

until 11 clusters remained. This took four more iterations. Five of the 11 samples of the output of Level 2 are shown in Figure 7. These will form the desired style templates. The 11 templates constructed from this procedure qualitatively look like they describe the 581 postmark samples in the dataset.

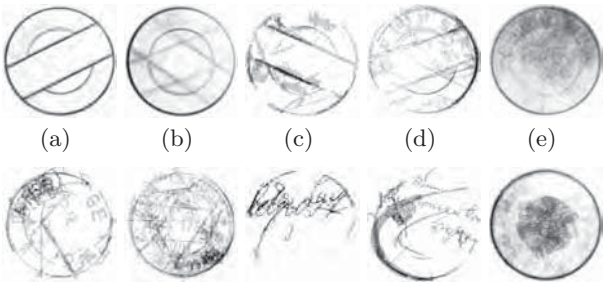
4.4 Comparative Single Stage Clustering

The multiple stage clustering procedure has the advantage over other clustering methods of reducing the noise in the process. For comparison the same base data was clustered using a single stage agglomerative clustering algorithm. A higher set distance, Eq. 1, threshold was used to produce a smaller number of clusters in a single step. Combining the data set to a small number of clusters in a single stage forces the clusters to be larger, which forces many non-similar symbols into the same cluster.

For the threshold chosen, a total of 34 clusters resulted after application of a single stage clustering algorithm. Ten of these are shown in Figure 8. While Figure 8a looks better than a similar output of the multistage approach, Figure 7a, the others Figure 8(b)-(j) are worse examples of category templates. Most of the other 24 results (not shown) have a similar appearance, where the structural information is not visible. Prominent features are averaged away, while noise remains visible. There was no mechanism to exclude outliers. If a higher threshold were used to produce in a single stage 11 instead of 34 clusters, some of these clusters would merge, but no additional structural information would result in the produced style exemplars. The multiple stage clustering approach allows structural information to be saved, while discounting the variable noisy data in these heavily degraded samples.

4.5 Image Retrieval

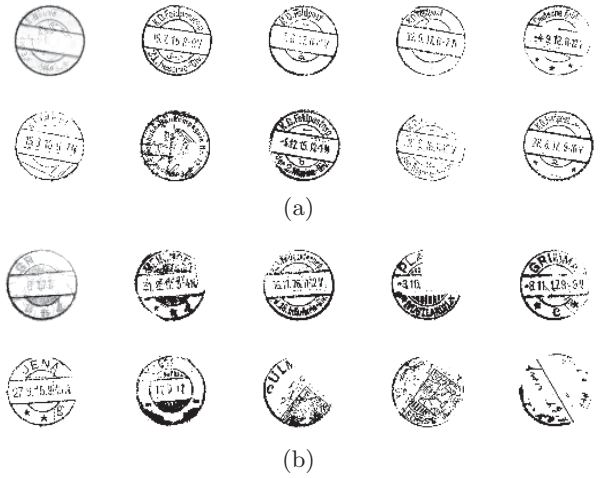
One use for which these template images is intended is to subdivide the postmark dataset for further processing. To show the feasibility of this, each of the 11 templates produced by the multistage clustering process proposed in this paper was compared to the 581 original postmark edge images. The template with the best match score (using Equation 2) was selected. Figure 9 shows nine of the results rotated to the angle of best match with the template shown



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 8: Output of single level clustering.



(a)

(b)

Figure 9: Examples of postmarks that best fit the templates (a) 2 and (b) 8. The first image is the template.

first. Template number 2 with its strong lines matches very well. Note that the lines in Template 8 are spaced differently than in Template 2 and the inner circle is partially filled. While these results are not perfect, they show preliminarily the descriptiveness of the resulting templates. Better retrieval methods exist and trying them with these templates as query images is potential future work.

5. CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

A cascade of unsupervised classifiers was used to form templates describing the structural forms of WWI feldpost postmark stamps. The use of multiple stages allowed averaging to reduce the noise, but not to wash away the details.

A simple, but effective, method was implemented to detect the postmarks in the full page image. It had a high success rate, comparable with others reported in the literature. Deeper analysis of the failures may increase that. Other ongoing work to separate the postcard content will make the postmarks easier to detect, as well as reduce occlusion.

The templates generated from the final stage of the cascaded clustering qualitatively look good. They show signifi-

cant improvement over the results of single stage clustering, that resulted in more templates. The templates are able to serve as query images for image retrieval to group the postmarks into reasonable clusters.

5.2 Future Work

It is difficult to match the images with one another when parts of the mark are missing. The penalty that occurs for the missing part, may exceed the score it received for the parts that match. Likewise when there is occlusion, the penalty for the occlusion should not exceed the reward for the match of the original. A more thorough investigation of matching metrics, such as used in line drawing research, might yield a better choice of metric. Improvements in the match metrics will increase the quality of the results.

In the future it would be beneficial to have features that were structural, as well as the current pixel features. A method that is robust to the fading, dropout and occlusion is needed. The text removal process could be replaced by one that requires components to be low curvature arcs and lines with a minimum length.

The postage stamps can be somewhat easily detected by examination of the background image estimates. They are large colorful or dark blobs on an otherwise mostly light image with uniform hue. Detecting these and adding that information to remove postage stamps from the postmark images would reduce the noise that they impart on this process. Work that has recently been done to separate the postmark from the underlying stamp image will allow semi-clean postmark samples instead of ignoring or blanking the area where the postage stamps were located. Reprocessing this dataset with the stamp images partially removed will show the net result. Similarly, as other work continues and the handwritten narrative text is identified to the point where it can be separated from the post marks, this process can be repeated. Literature has often suggested that the document recognition process should include a feedback approach to incorporate results from future steps. The identification of the postmark templates will allow them to be identified and removed, thus making the background text more accessible, which will make the postmarks clearer to work with.

Applying this technique to the complete postcard collection will better show its capabilities. Trying other base clustering algorithms that are not as computationally intensive as heirarchical clustering will better suit processing of larger collections.

Trying this method on other datasets would also be good for comparison. The datasets from Llados et al. [**?**] was considered, but the images are all very clean and thus a single level clustering algorithm works well on them. Trying to artificially degrade those images to remove parts of them and make parts faint or have larger quantities of overwritten material might make this dataset useful for comparison.

Once the structure of the postmarks has been identified, they can be either matched to the original image, through a retrieval method such as in Section 4.5. The postmarks that match a particular form can be further clustered to find the sub categories and form exemplars that include frequently occurring text content. A full taxonomy of postmark styles used during this period could then be produced.

6. ACKNOWLEDGMENTS

This work was partially funded by a grant from the DAAD *Deutscher Akademischer Austausch Dienst*.

The authors would also like to thank Dr. Britta Bley, Dortmund, Germany, for providing the collection of historical postcards.

7. REFERENCES

[1] Europeana 1914–1918 — untold stories & official histories of WW1. http://www.europeana1914-1918.eu.

[2] B. Bley. Feldpostkarten im 1. Weltkrieg (Feldpost Postcards of World War I). Private Collection.

[3] C. Brocks. *Die bunte Welt des Krieges: Bildpostkarten aus dem Ersten Weltkrieg 1914–1918 (The Colorful World of the War: Picture Postcards from the First World War 1914–1918)*. Klartext-Verlag, Essen, 2008. (in German).

[4] J. Chen, M. K. Leung, and Y. Gao. Noisy logo recognition using line segment Hausdorff distance. *Pattern recognition*, 36(4):943–955, 2003.

[5] D. Doermann, E. Rivlin, and I. Weiss. Applying algebraic and differential invariants for logo recognition. *Machine Vision and Applications*, 9(2):73–86, 1996.

[6] X. Dong, J. Dong, and S. Wang. Segmentation of Chinese postal envelope images for address block location. In *LNCS Advances in Visual Computing*, volume 5876, pages 558–567. Springer, 2009.

[7] L. F. Eiterer, J. Facon, and D. Menoti. Postal envelope address block location by fractal-based approach. In *Proceedings 17th Brazilian Symposium on Computer Graphics and Image Processing*, pages 90–97, Brazil, October 2004.

[8] G. A. Fink, L. Rothacker, and R. Grzeszick. Grouping historical postcards using query-by-example word spotting. In *Proc. Int. Conf. on Frontiers in Handwriting Recognition*, Crete, Greece, 2014.

[9] M. Gori, M. Maggini, S. Marinai, J. Sheng, and G. Soda. Edge-backpropagation for noisy logo recognition. *Pattern Recognition*, 36(1):103–110, 2003.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[11] B. Lamiroy and Y. Guebbas. Robust and precise circular arc detection. In *8th International Workshop on Graphics Recognition. Achievements, Challenges, and Evolution, GREC 2009*, volume 6020 of *Lecture Notes in Computer Science*, pages 49–60, La Rochelle, France, July 2010. Springer-Verlag.

[12] Y.-J. Liu and F.-C. You. Postmark date recognition based on machine vision. *Physics Procedia*, 33:819–826, 2012.

[13] H.-L. Peng and S.-Y. Chen. Trademark shape recognition using closed contours. *Pattern Recognition Letters*, 18(8):791–803, 1997.

[14] P. Petej and S. Gotovac. Comparison of stamp classification using SVM and random ferns. In *IEEE Symposium on Computer and Communications (ISCC)*, pages 850–854, Split, Croatia, July 2013.

[15] C. C. Reyes-Aldasoro. A retrospective shading correction algorithm based on signal envelope estimation. *Electronic Letters*, 45(9):454, 2009.

[16] K. Roy, S. Vajda, U. Pal, B. B. Chaudhuri, and A. Belaïd. A system for Indian postal automation. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 1060–1064, 2005.

[17] P. P. Roy, U. Pal, and J. Lladós. Document seal detection using ght and character proximity graphs. *Pattern Recognition*, 44(6):1282–1295, June 2011.

[18] H. Yuan, H. Ma, and X. Huang. Image-based stamp extraction for enhanced postal automation. In *Image and Signal Processing, 2008. CISP'08. Congress on*, volume 3, pages 672–676, 2008.

[19] G. Zhu and D. Doermann. Automatic document logo detection. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 864–868, 2007.

[20] G. Zhu, S. Jaeger, and D. Doermann. A robust stamp detection framework on degraded documents. In *Proc. SPIE 6067, Document Recognition and Retrieval XIII*, volume 6067, page 60670B, San Jose, CA, January 2006.