The Creation of Synthetic Digital Ground-Truth Images of Historic Cosmic Ray Data Recordings

Vincent Mattana, Günther Drevin(✉), and Pierre Roux

North-West University, Potchefstroom Campus, Potchefstroom, South Africa
gunther.drevin@nwu.ac.za

Abstract. The aim of this paper is to develop a set of algorithms for the automated construction of synthetic digital ground truth images from historic cosmic ray recordings. These images can subsequently be used to test data extraction algorithms. This takes place in a larger research context of an effort to retrieve and digitize the data contained within more than 20 years (1934–1956) of historic cosmic ray data from around the world. The creation of synthetic ground truth images can logically be broken down into component tasks, which can be approached individually. These tasks include: binarization, segmentation, as well as generation of optical artefacts and distortions. The approach and details of the algorithm are described.

Keywords: Image processing · Cosmic rays · Segmentation · Digitization · Degrade · Document images · Local adaptive binarization · Soft decisions

1 Introduction

The aim of this project is to develop a set of algorithms to automatically construct synthetic ground-truth images from historic cosmic ray recordings which can be used to compare the accuracy and efficiency of different interpretation and data extraction algorithms. When designing a data extraction algorithm, ground truth images are required to score the algorithm's performance and to compare it to other algorithms. These synthetic ground truth images are critical in rating the accuracy of the algorithms. In a sense this is the first step towards the complete extraction and interpretation of historic cosmic ray recordings. The future steps could include data detection, extraction of the numeric values represented by these data recordings, as well as interpretation of these results (which document the ionization levels in a cosmic ray detector). Eventually the cosmic ray data contained within these recordings could be released for further study. However, that is beyond the scope of this project, in which the goal is to create a synthetic ground truth image, which replicates the condition and appearance of the recorded data, as well as having a benchmark against which to test competing algorithms in the following steps of the data rescue effort. The synthetic image must however contain some degree of distortion, so that

B. Lamiroy and R. Dueire Lins (Eds.): GREC 2015, LNCS 9657, pp. 19–30, 2017.
DOI: 10.1007/978-3-319-52159-6_2

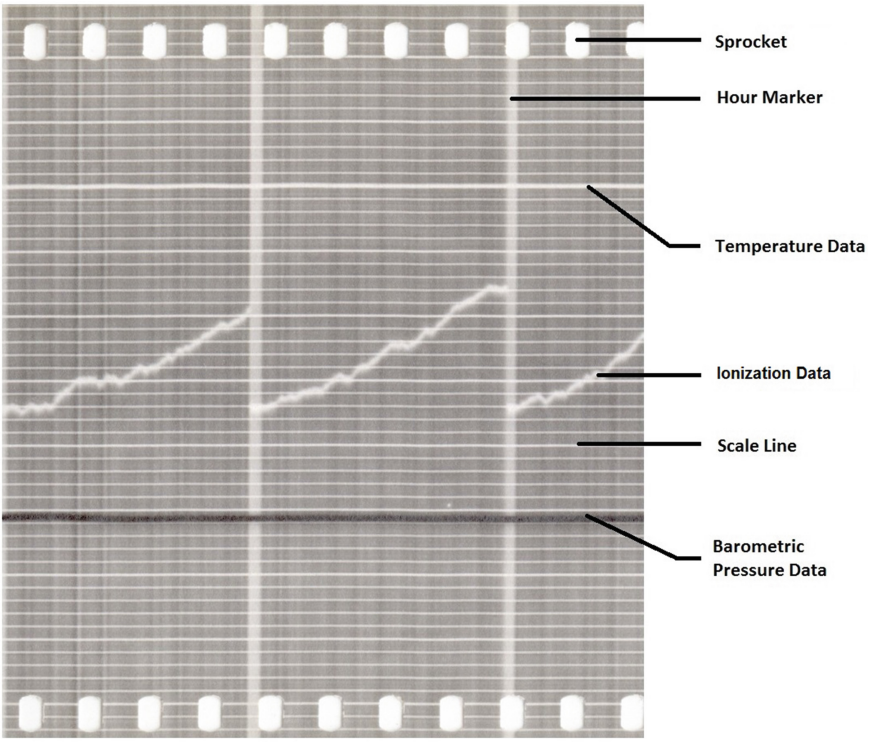

Fig. 1. Diagram of an original section of the photographic paper.

the tested algorithm can function in an environment as similar to the original image as possible [3]. This is because the more similar the synthetic image is to the original, the more relevant the results of the testing of the algorithms will be. Once the ground truth and synthetic images are complete and verified, the algorithms responsible for their creation can potentially be integrated into the data extraction and interpretation algorithms, to provide accuracy statistics on the results.

1.1 The Model C Detector

The Carnegie Institute funded a project to develop a precision cosmic ray recorder, which was undertaken by Compton *et al.* [1]. A number of these devices, the Model C cosmic ray ionization chamber (henceforth the Model C detector), were manufactured by the team [1]. Each of them recorded cosmic ray data on photographic paper.

The true value of these recordings lies in the massive collection of data that was generated from the manufacture of the recording devices, in 1934, up to 1956 when neutron monitors began to replace the Model C detectors.

More than a 100 station-years of recordings were generated by the Model C detector. The structure of the images on these photographic papers can be described by investigating their components as well as the recording process: Each image consists of horizontal scale lines and vertical hour and day markers, Fig. 1. Additionally, there are data recordings for cosmic ray intensity, barometric pressure, and temperature. The cosmic ray intensity was recorded continuously by recording the electric charge accumulating in the ionization chamber, Fig. 2. The recording process involves the mounting of the recording paper onto a sprocket drive, aligning the scale grid, as well the recording needles. It is assumed that adjustments were not performed perfectly every time due to human error, and as such could have led to distortions within the recordings, which will be discussed later. The contrast and quality of the images also vary due to inconsistencies in the chemical processing of the photographic paper. Other examples of distortions include hand written annotations, punched holes, and stickers. Furthermore bleed-through of annotations on the back of the photographic paper are also present.

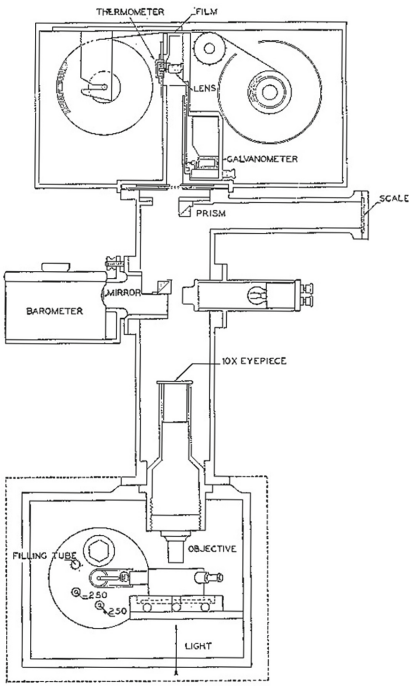

Fig. 2. Detailed drawing of electrometer box, optical system, barometer, and recording camera. Adapted from [1].

2 Methodology

The method used for this study can be described as the hypothetico-deductive method [2]. Our approach followed a step-wise improvement of the source image, and is described by the following steps:

1. Increasing the contrast of an original image.
2. Binarization of the resulting image.
3. Segmentation of the image to extract components of the image such as hour markers, scale lines, data lines and sprockets.
4. Production of an approximate ground truth image, using the segmented image.
5. Mapping of distortions and optical artefacts found in the original images onto the ground truth image, to produce a synthetic image.

6. Creation of a textured synthetic image, using a texture only image created from an original image.
7. Comparison of the synthetic image to the original image.

The photographic paper that the image is recorded on can contribute to the graininess of the image, and any imperfections of the material used to record the original document will be carried over into the digital format. The process of scanning these photographic papers was undertaken by a private contractor. The images measure approximately 6000 by 905 pixels, have a pixel depth of 24 bits, and on average document a 19 h period.

2.1 Contrast Correction

The first step in creating ground truth data is to perform contrast correction on the source image. The generation of a ground truth data image is simplified by using a single high quality source image such as Christchurch 1937-08-02 Fig. 3. This image was selected due to its good condition, and each element of the recording was immediately recognizable by inspection. Global histogram equalization was used to improve the contrast of this image.

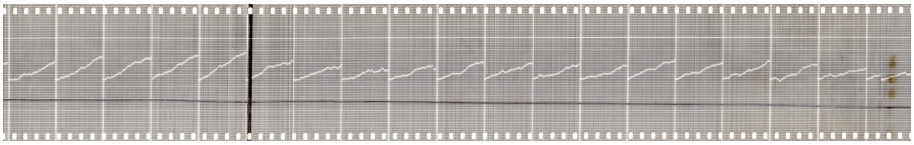

Fig. 3. The single high quality source image used (Christchurch 1937-08-02).

2.2 Binarization

The process of preparing a document image for the extraction of text from the document is known as document image binarization, however the intermediate steps of this method has much potential for segmenting the cosmic ray recordings. Many different techniques for document image binarization have been proposed, and could be used to binarize the historic cosmic ray recordings, but the work of Sauvola and Pietikainen [11] and the improvements made thereupon by Gatos *et al.* [5] were the most effective in our application (Fig. 4). Otsu's method [8] was also used to isolate the hour markers. These techniques have all the necessary methods we need for the binarization of the historic cosmic ray recordings. Another option is adaptive thresholding, which selects different threshold values for different regions, by inspecting the gray-level intensities in a mask across the image. A number of different adaptive thresholding algorithms exist, such as Niblack's algorithm [6], which is built upon by Sauvola and Pietikainen [11], and then further refined by Gatos *et al.* [5], by using an adaptive Weiner filter to remove noise.

A difficulty in adaptive binarization is that a decision has to be made on the size of the mask to be used. The mask must be big enough to ensure that a sufficient selection of background (texture) pixels are captured, while maintaining a small enough footprint as to prevent the averaging of background pixels over nonuniform intensities. To counteract this, domain based information is used to check that the results give expected values [7]. It is essential to remove any noise or artefacts in an image before binarization since this can become detrimental in the later steps. If the image contains any texture, it is important to do contrast enhancement so that the foreground elements become more apparent from the background. This is achieved by using an adaptive Weiner filter, which uses statistics from local neighbouring pixels.

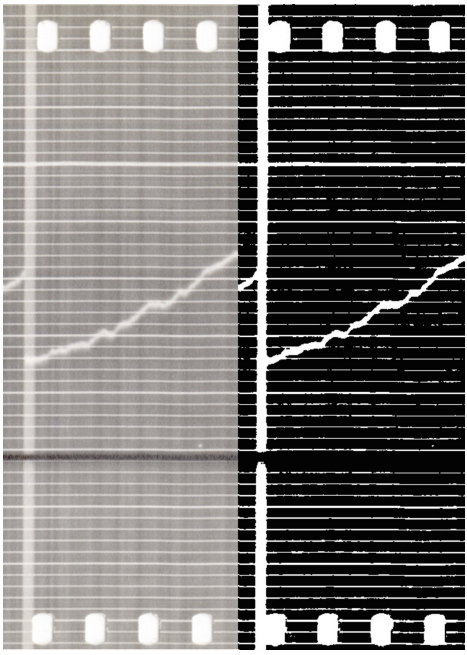

Fig. 4. The result of binarizing the image on the left.

2.3 Segmentation

Document image processing is concerned with all aspects of "working with" documents, from scanning through all the pre-processing steps to the final information extraction and recognition. Many algorithms exist within this field to extract text from documents, using segmentation, which can be described as the division of the image into logically distinct segments. These algorithms can be adapted to fit the needs of our objective, namely the creation of a ground truth image, which will eventually be distorted from its ideal condition to produce a synthetic image, which is comparable to the original image. The segmentation of the image is an iterative process, eliminating certain elements in an effort to isolate a specific attribute. Once an attribute of the ground truth image has been extracted, the algorithm reverts to working from an earlier version of the image, usually the binarized image. The image is then processed again to yield only the desired attribute. A number of different techniques were used, depending on the properties of the attribute to be extracted. The attributes that were extracted are shown in Fig. 1, and will be discussed now.

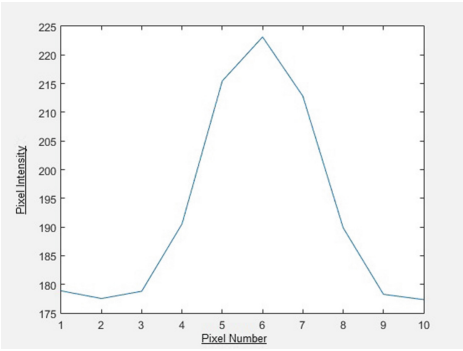

Fig. 5. Cross section of a scale line.

The Scale Lines. Scale lines run horizontally along the photographic paper, varying slightly over the length of the recording. Exploiting this directionality simply requires using a horizontal Prewitt filter [10] to only detect horizontal edges. The scale lines are represented by double edges on the resulting image. These double edges are filled by using morphological filtering in the vertical direction. The result is a line a few pixels thicker than the original. The scale lines however are still disrupted by the edges of intervening data lines. This effect is removed by using horizontal morphological filtering. The result is an image of thinner scale lines, disrupted only by the position of sprockets, temperature data, cosmic ray data, barometric pressure data and hour lines. To proceed further one must determine the number of the scale lines, as well as their positions. The height of the image, divided by the number of scale lines results in the number of pixels that should ideally be between successive scale lines. This information, along with the positions, are used to generate an ideal image of approximated scale lines. With the use of an ideal scale line image, it becomes a simple task to obtain samples of data for each individual scale line by following a line on the ideal model of the scale lines, with a seek range of up to half the number of pixels between each scale line. These samples are in turn used in regression analysis to determine a best fit linear function for the given data. The regression of the scale lines is based on the original image, and as such the ground truth scale lines are accurate representations thereof. A cross section of a typical scale line is shown in Fig. 5.

The Hour Markers. The same approach used to extract the scale lines will not work on the hour markers. Even if the Prewitt filters [10] are rotated 90°, there will still be false hour markers generated, since the majority of historic cosmic ray recordings contain spurious vertical edges. This is mainly attributed to the motion of the photographic paper moving through the recording device, which is not precisely constant nor uniform. A different approach was taken: By using the disruptions in the scale lines (in the binarized image of scale lines) the hour marker positions are clearly visible as gaps in the scale lines. This process is automated by using a parameter defining a threshold for locating the hour markers from gaps in the scale lines. Once these position are identified, ideal vertical lines are inserted at these locations. The hour markers are approximated using the average-pixels-between method and are then thinned to produce the ground truth hour markers. This effect is dependent on scale lines being appropriately located. Bearing in mind that the recording images each document a period of 19 h, a neighbourhood around the detected hour markers was inspected for any

other detected hour markings, which could be false positives, as hour markers should appear approximately every 285 pixels, as an hour is approximately 300 pixels. Each hour marker has a width of approximately 15 pixels which represents 3 min of recording time, during which the ionization chamber was grounded.

The Sprockets. The tiny holes at the top and bottom edge of the photographic paper, are referred to as sprockets, and are used to advance the photographic paper. These sprockets are shaped like a flattened circle. The sprockets are located in a binarized, and smoothed image, where only the faintest remains of the sprockets are present. From this image it is possible to locate the sprocket "seed" positions, and construct an ideal sprocket at the location of each "seed" pixel. To simulate the shading found in the original image, a copy of the ideal sprocket image was used, and buffered (shifted 3 pixels towards the centre). This buffered image was then subtracted from the original, and a motion blur filter was applied to roughen the edges. A sample sprocket image and cross section of a typical sprocket hole is shown in Fig. 6.

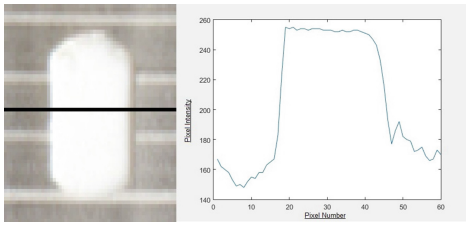

Fig. 6. A sample sprocket image and cross section of a typical sprocket hole.

Data Lines. There are 3 types of data recorded on the image. The thinner horizontal line represents the temperature data, the discontinuous line represents the cosmic ray ionization values, and the thicker, dark, horizontal line represents

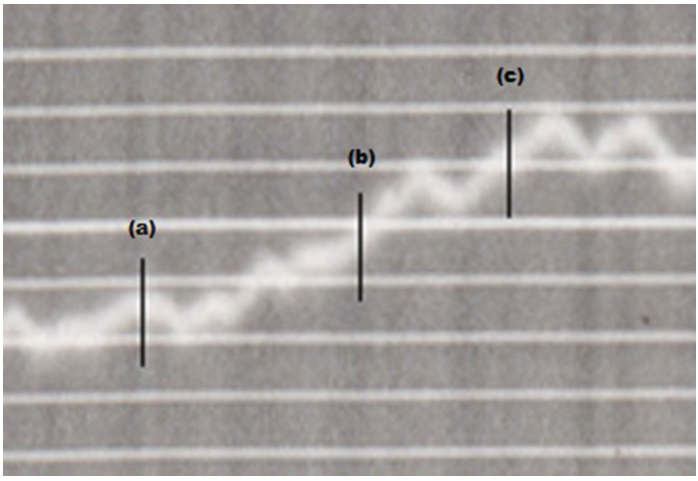

Fig. 7. Cosmic ray data line cross sections taken along the lines.

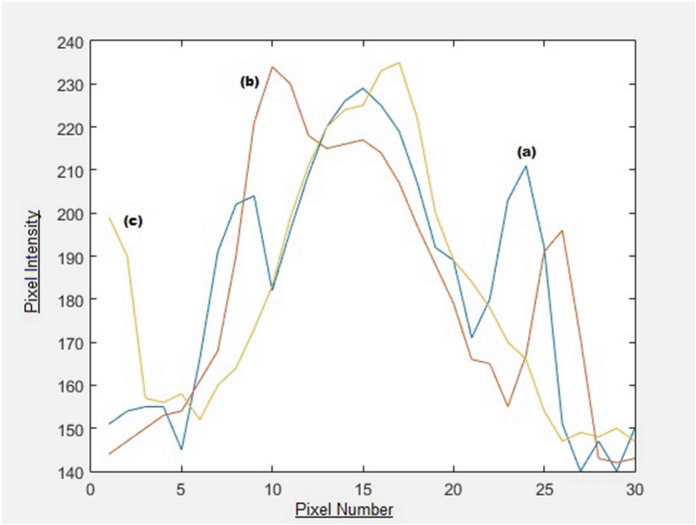

Fig. 8. Cross section of the cosmic ray data line, taken at varying points.

the barometric pressure. Ideally, the data lines would be extracted automatically, however, doing so reliably is difficult on some images. The data was instead extracted using a cursor to select data lines. This lets one determine the centre of the line, giving an indication of where the data lines can be found on the original image. This forms the basis for the data lines in the ground truth image. Cross sections of the data line at varying points are shown in Figs. 7 and 8.

2.4 Production of Ground-Truth Images

The final ground truth image is produced by adding together all the attribute images obtained while performing segmentation. This ground truth image represents the core elements of the original image. As this is a ground truth image, there is no texture, nor distortions included. The texture, and distortions will be discussed in the next section. The result is shown in Fig. 9.

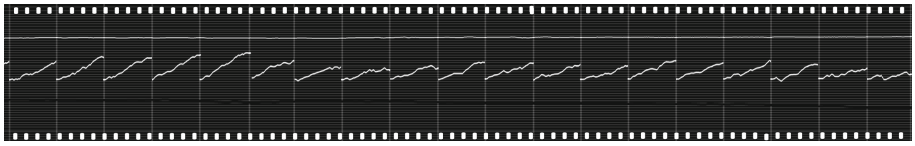

Fig. 9. The result of combining the composite segmented images.

2.5 Distortions

At this stage, the ground truth image has been produced with single pixel thick scale lines and data lines, as well as accurate hour markers, and sprockets. However, one will notice that this ground truth image does not resemble the original image, other than as a representation of the data elements contained in the original image. Thus the ground truth image has to be modified by some degree to represent the source image's texture and distortions. This is accomplished in a modular fashion, repeating similar algorithms for each component of the image.

Two main features of distortion are prominent in the hour markers, namely: edge sharpness and rotation [4]. Hour markers are produced on the photographic paper by deactivating the lamp for three minutes every hour. The orientation of the slit, however, is not always perfectly vertical. The angular rotation of the hour markers is found using the angle between the regressed scale lines and the ideal hour markers. These hour markers overlap with scale lines, and as such, the scale line intensities are subtracted before any smoothing is applied. The scale lines were distorted using a Gaussian filter that simulates the interference pattern of light, when passed through a narrow slit.

The majority of each sprocket is undistorted. However, one can see shadows cast by the photographic paper when comparing ideal sprockets to the sprockets on the original image (these distortions are mainly due to a lack of light and is perceived as a shadow of the image). Such shadows are added to ideal sprockets by adding a gradient filter over the ideal sprocket. It can be noted that the sprockets are the largest of the artefacts, as they affect the photographic film itself. This allows us to assume that there is no useful information to be gained in the sprocket areas, and as such can be disregarded in future algorithmic steps.

The data lines on the photographic paper are blurred, and do not have sharp edges. The distortion algorithm mimics this imprecision with a collection of image processing techniques tailored to simulating the histogram of the corresponding data line. This process results in a distorted ideal data line nearly identical to the source image's data line.

Reproducing Optical Artefacts. When considering the effects that occur regarding light and photographic paper, it is important to remember two things. The first thing to remember is that the real world is continuous, and not discrete like the digital world. The second thing to bear in mind is that when two waves superimpose, the resultant wave has either greater or smaller amplitude than the source waves. This physical phenomenon is referred to as interference, and it can cause ringing, an effect which causes dilation of point like illumination with bands of bright and dark. Usually these bands are so close to each other, and also moving, that they are blurred on the photographic paper. This blurring and ringing results in a broader and less defined data line. Thus it is necessary to generate such a distortion for the synthetic image. Various methods exist to achieve the desired results, and the fastest computational method was selected: By subtracting a slightly blurred image from a heavily blurred image, the results show a thin "ringing" around the original object. When this image and the

distortions of all the attributes are added together, the resulting image contains optical artefacts, resembling the original distortions closely.

2.6 Texture

The texturing process is accomplished through a number of steps. The first step is to replace all the attributes' pixels with pixels from nearby texture regions. This image is then smoothed, and a histogram equalized version of the image is subtracted from this smoothed image, resulting in a mid-tone texture only image. From this point, the mid-tone image is binarized, once with a low threshold, and again with a high threshold. These binarized images represent the highlights, and darker regions of the image, respectively. Both of these binarized images are then added to the mid-tone image to yield the texture only image. This texture-only image is then combined with the distorted ground truth image to produce the result seen in Fig. 10.

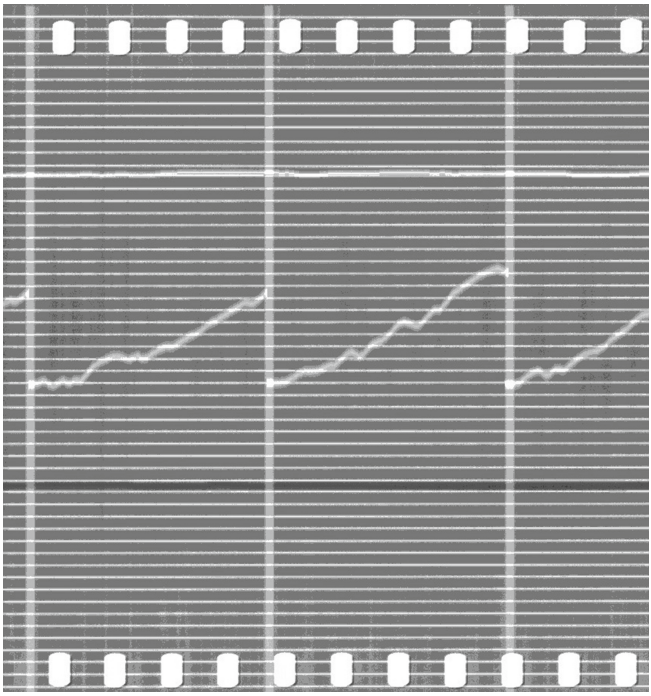

Fig. 10. The result of adding the texture to the distorted ground truth image, yielding the synthetic image.

3 Results

To adequately find a measure of success, the accuracy of the ground truth attributes was determined. The algorithm was tested by creating synthetic images from images of differing quality: Good, average and poor. These are each a single recording from: Christchurch (1937-08-02), Fredericksburg (1958-08-02) and Christchurch (1945-11-03), respectively. These historic recordings all have a similar recording format in common, in that all attributes have the same colour/intensity relative to their background. This is important to note, as some images have black hour markers, or other discrepancies. Comparison of image transformation algorithms can be done empirically by measuring the Mean Square Error (MSE) of each image [9]. However this will not suffice for the task at hand, since the synthetic generation of a texture will differ greatly in pixel intensities at specific coordinates which will in turn influence the MSE. The texture might be visually very similar to that of the source image, but the same result will not be represented by its MSE. Because of this, the texture and colour/intensity distortions are very subjective attributes to compare. Attributes such as scale lines, hour markers and sprockets will have their measure of success determined by the number of attributes matching the source image. To compare the results more clearly, the synthesised image was overlaid on the original image. These attributes are then coloured, and the attributes where compared. The results of this comparison are shown in Table 1.

Table 1. Empirical analysis of results

Measure	CHE ('37-08-02)	FRE ('58-08-02)	CHE ('45-11-03)	Average Accuracy
Scale lines	54/56	53/56	51/56	94.05 %
Hour markers	20/20	13/20	8/20	68.33 %
Sprockets	140/144	142/144	32/144	72.69 %
Texture	8/10	7/10	4/10	63.33 %
Data line consistency	19/20	13/20	15/20	78.33 %
Accuracy	93.73 %	78.65 %	53.66 %	75.35 %

4 Conclusion and Further Work

The aim of this study was to create a digital synthetic ground truth image from existing historic cosmic ray records. The goal of creating an image similar to the source image, while also containing the ground truth data, has been sufficiently achieved. The algorithms perform adequately under ideal circumstances, however, it preforms less effectively when applied to low quality source images. Improvements can be made to the texturing process, the automated thresholding

for binarization, the data line and hour marker extraction, as well as compensating for the inherent angling (or tilt) in some of the hour markers and scale lines.

References

1. Compton, A.H., Wollan, E.O., Bennett, R.D.: A precision recording cosmic-ray meter. Rev. Sci. Instrum. **5**, 415–422 (1934)
2. Dodig-Crnkovic, G.: Scientific methods in computer science. In: Proceedings of the Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden, Skövde, Suecia, pp. 126–130 (2002)
3. Drevin, G.R.: Adaptive frequency domain filtering of legacy cosmic ray recordings. In: Proceedings of the 11th Joint Conference on Information Sciences, pp. 15–20 (2008)
4. Drevin, G.R., Moraal, H., McCracken, K.G.: Determining the skew and scale in images of Compton-Wollan-Bennett ionization chamber recordings. Information Sciences 2007, pp. 888–894. World Scientific, Singapore (2007)
5. Gatos, B., Praktikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recogn. **39**, 317–327 (2006)
6. Niblack, W.: An Introduction to Digital Image Processing. Prentice-Hall, Upper Saddle River (1985)
7. O'Gorman, L., Katsuri, R.: Document Image Analysis. IEEE, New York (1997)
8. Otsu, N.: A thresholding selection method from gray-level histograms. Automatica **11**(285–296), 23–27 (1975)
9. Ponomarenko, N., Krivenko, S., Egiazarian, K., Astola, J., Lukin, V.: Weighted MSE based metrics for characterization of visual quality of image denoising methods. In: Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2014) (2014)
10. Prewitt, J.M.S.: Object enhancement and extraction. Picture Process. Psychopictorics **10**(1), 15–19 (1970)
11. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recogn. **33**, 225–236 (2000)

http://www.springer.com/978-3-319-52158-9

Graphic Recognition. Current Trends and Challenges
11th International Workshop, GREC 2015, Nancy, France, August 22–23, 2015, Revised Selected Papers
Lamiroy, B.; Dueire Lins, R. (Eds.)
2017, X, 149 p. 83 illus., Softcover
ISBN: 978-3-319-52158-9