

3. Experimental process

3.1. Introduction

This chapter describes the experimental process of designing the methods applied to the data as well as the data used during experimentation.

The goal of this study is to create a method that is capable of adaptively binarizing and extracting data from images of cosmic ray recordings, without any additional user input. Additional image markings such as hour markers and scale lines are not targeted by this process as they may be more accurately binarized using methods mentioned by Du Plessis (2010), which are applied in the frequency domain.

3.2. Test data

The images that were used during the experimental phase of this study were chosen from the output of four recording stations which implemented the Model C ionization chamber.

The locations of these stations are:

- **Huancayo - Located in Peru at: Latitude 12.0°S and longitude 75.3°W , 3350 meters above sea level;**
- **Christchurch - Located in New Zealand at: Latitude 43.5°S and longitude 172.6°E , 8 meters above sea level;**
- **Cheltenham - Located in the United States of America at: Latitude 83.7°N and longitude 76.8°W , 72 meters above sea level;**

3. *Experimental process*

- Godhavn - Located in Greenland at: Latitude 69.2°N and longitude 53.5°W, 9 meters above sea level;

Samples of data produced by these stations were selected to be used during the experimental phase of this study. Each image contains roughly fifteen hours of data, after each hour the recording device was reset and the origin of the data line of the next hour was shifted back to a starting position that remained constant throughout the image.

Each image contains data lines that need to be extracted and binarized. Several unwanted image objects are also present in the image that can reduce the accuracy of the binarization process. These objects are:

- Sprocket holes: The holes in the original photographic strip that were used to feed the strip into the recording device. These holes are present in the digital image as white rounded rectangles;
- Scale lines: The scale lines within the image can affect the accuracy of the process. They have similar intensity values to the data lines and must not be extracted along with the data lines;
- Hour markers: These thick lines spanning the height of the image can negatively affect connectivity between data line segments;
- Empty space: The empty white spaces at the top and bottom of each data image;
- Temperature lines: Although these lines are also extracted during the binarization process, they need to be removed from the image while extracting the data line;

The images used during experimentation are:

- Image A: This image was recorded at the Cheltenham station. This image has a very low average intensity value. The data within the image has such low intensity levels that it is difficult to separate it from the background. Hour markers and scale lines are not visible in this image. The image can be seen in Figure 3.1;

3. *Experimental process*

- Image B: This image was recorded at the Cheltenham station. This image also has a low average intensity value, but the data in this image is mostly visible along with some scale lines and hour markers. The image can be seen in Figure 3.2;
- Image C: This image was recorded at the Cheltenham station. This is a relatively well recorded image, where the intensity values of the background and foreground differ enough to easily differentiate between them. Scale lines, hour markers and temperature data are also clearly visible. The relatively straight shape of the data line may pose a problem when the process attempts to remove the horizontal line of temperature data from the image. The image can be seen in Figure 3.3;
- Image D: This image was recorded at the Godhavn station. Although the background and foreground intensities differ enough to make the data clearly visible, the contrast of the image is high enough that the intersections between the scale lines and data lines affect the accuracy of the extracted data. The high intensity values of the hour markers may also affect accuracy. The image can be seen in Figure 3.4;
- Image E: This image was recorded at the Huancayo station. It has the same contrast characteristics as image C, but the hour markers have very low intensity levels and temperature data is not visible. The image can be seen in Figure 3.5;
- Image F: This image was recorded at the Christchurch station. The image has a very high average intensity value. Several parts of the background have the same intensity as the data lines. Temperature data is not clearly visible within the bright image. The image can be seen in Figure 3.6;

All these images also have non-uniform background intensities and the data lines within the darker images also have varying intensity levels.

These images contain some of the most degraded data lines of the entire original data set, with the data lines still being relatively visible. The images also represent the variety of challenges presented by the data. Each image contains several of these challenges with varying degrees of intensity. If the process can successfully extract binarized data lines from these images, it will be able

3. *Experimental process*

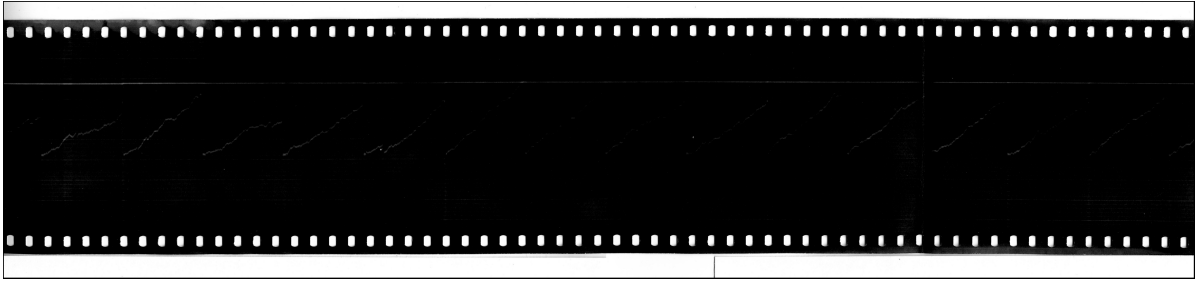


Figure 3.1.: The original image A.

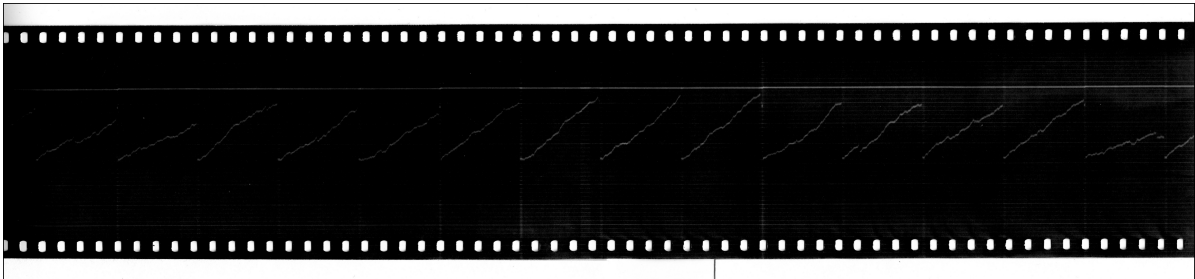


Figure 3.2.: The original image B.

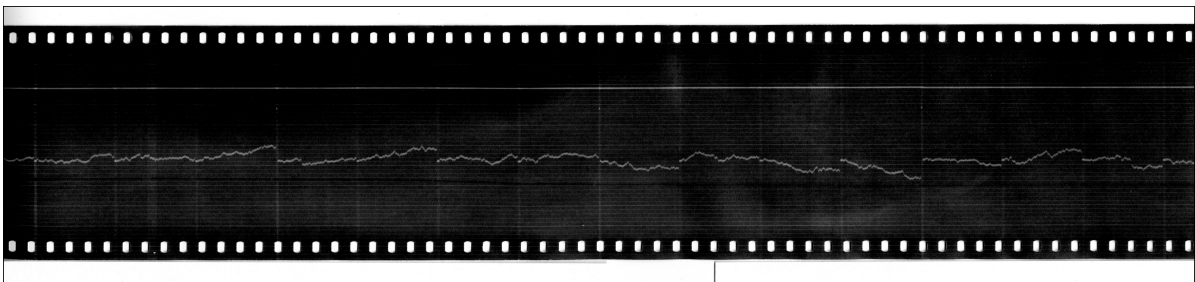


Figure 3.3.: The original image C.

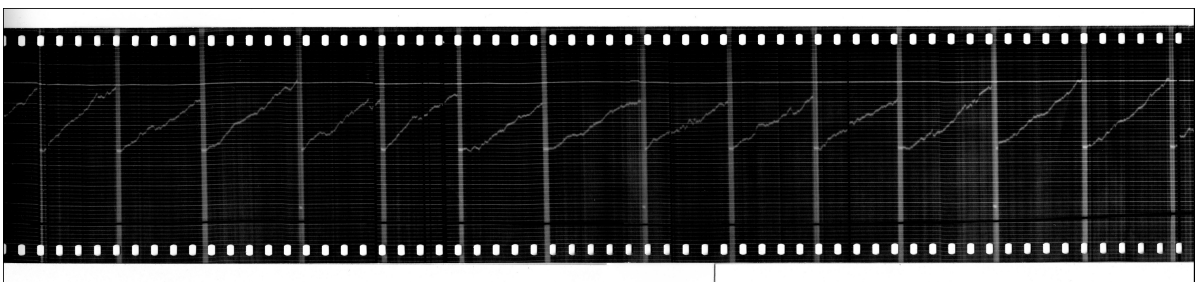


Figure 3.4.: The original image D.

3. *Experimental process*

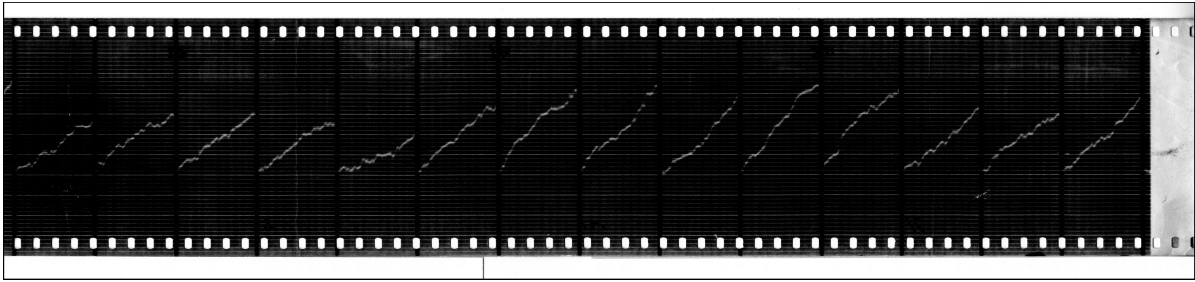


Figure 3.5.: The original image E.

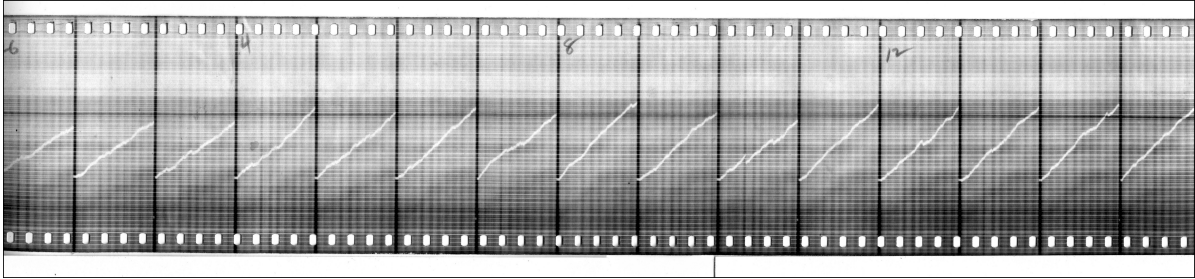


Figure 3.6.: The original image F.

to binarize any image from the original dataset that contains enough identifiable data line pixels to reconstruct from.

Other images from the original data set that contain more visible data lines would not be as difficult to binarize, and thus would not help to test the capabilities of this adaptive process.

From this point onward these images will be referred to by their alphabetical designation.

The resolution of the images is 7006 x 1622 pixels and is not modified at any point during the binarization process. The large original images will take more time to process, but this approach promises the best accuracy. During the course of this research, accuracy is a higher priority than processing speed.

3.3. Design Process

Early in the experimental phase it was discovered that most popular document binarization methods could not be applied to the images mentioned above to extract data from cosmic ray images. The non-uniform background and data line intensities resulted in the popular global and local binarization methods, which are mentioned in Chapter 2, being unsuccessful, as shown in Figure 3.7.

The images within Figure 3.7 can be described as follows:

- A: Segment of the original Test image C;
- B: Application of a global threshold of 220;
- C: Application of a global threshold of 240;
- D: Application of Eq. 2.34 with a 7 x 7 window;
- E: Application of Eq. 2.34 with a 11 x 11 window;
- F: Segment of the original Test image F;
- G: Application of a global threshold of 60;
- H: Application of a global threshold of 80;
- I: Application of Eq. 2.34 with a 7 x 7 window;
- J: Application of Eq. 2.34 with a 11 x 11 window;

With the application of a global threshold, all pixels with higher values than the threshold are assigned a value of 255 (white) and all other pixels are assigned a value of 0 (black).

3. Experimental process

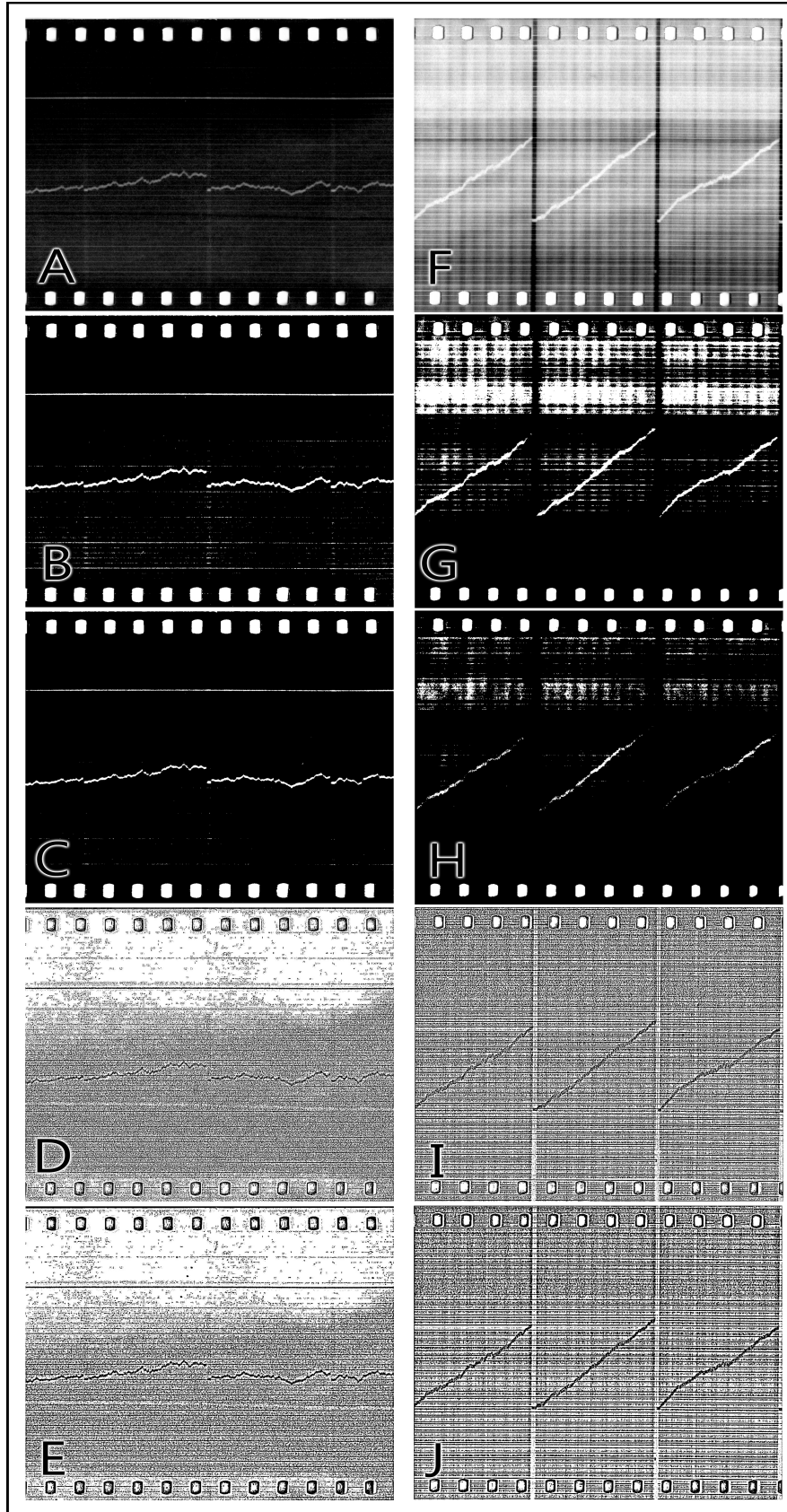


Figure 3.7.: Results of applying existing binarization methods to cosmic ray data.

3. Experimental process

With the application of Eq. 2.34 (Sauvola and Pietikäinen's foreground extraction equation), the results are inverted so the data lines, which are originally white appear as black data lines. Eq. 2.34 is a crucial step in the initial foreground estimation in both Sauvola and Pietikäinen's binarization method and the binarization method of Gatos *et al.*. If that step cannot be applied to the cosmic ray images, then those methods would not be successful when applied to the same images.

This made it clear that new methods needed to be created specifically for this type of image. All the methods and techniques discussed in the literature study were considered during the design process.

The iterative experimental process was conducted as follows:

1. Design method: Methods are designed to be applied to the image in order to achieve a specific result, such as removing noise or image objects. Document specific knowledge is applied so these methods target specific parts/characteristics of the image;
2. Apply the method to the image: The method is tested on the set of data images to see how the different image characteristics affect the result;
3. Examine results: The output image of the method is examined to determine if any variables should be increased/decreased or if an entirely new approach to the problem should be taken;
4. Repeat: The altered method is applied to the image to determine how the new variables affect the result. It is important to carefully study the results of each test image as the process of altering a variable to improve the results of one image may negatively affect the results of a second image, by increasing the effect of an unwanted object within that second image;
5. Continue: Once the method produces satisfactory results for all test images, then that output image is used as the input image of the next method to be designed;

3. Experimental process

Many of the methods described in this chapter could be combined to decrease processing time, but the methods are kept separate to allow the results of each step in the process to be inspected carefully so the effect of any changes made to the process will be clearly visible in the resulting images.

The application of document specific knowledge is crucial during this design process. For example, the fact that the data line within the image is usually the highest intensity object in the image is the core concept of the entire data extraction process.

3.4. Success

The success of the process is confirmed visually. If the resulting image is accurately fitted over the original image, then the fewer data pixels and thinner data lines within the binarized image should fall within the thicker/larger data areas of the original image.

The only data source to compare the results against, to establish the success of the process, is the original image itself. While a visual evaluation of the process results may seem subjective, the only alternative comparison would be against a manually hand drawn version of the original data lines. The process of drawing these lines would rely on the same visual analysis of the original. Thus, such a comparison would be no less subjective than comparing the process results to the original image itself.

3.5. Technology/Software

Most of the experimental work was done using MATLAB R2012a. MATLAB is a high-level language that is ideal for coding mathematical formulas. Graphical representation of results, effortless variable tracking and control and the ability to test and edit algorithms and parts thereof without needing to compile the code makes MATLAB the ideal software for use during these experiments. MATLAB handles matrix manipulation very well which has made it the obvious choice of testing platform for this study.

The next chapter describes the data extraction and binarization process in detail.