Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master's Degree in ICT for Internet and Multimedia

Automated Detection of Irregularities on Magnetic Audio Tapes Using Frame Differencing

Supervisor:
Prof. Sergio Canazza

Co-Supervisor:
Alessandro Russo
Matteo Spanio

Candidate:
Zafer Çinar
2041428

Academic Year 2023-2024

Abstract

Preserving the integrity and authenticity of magnetic audio tape recordings involves identifying any changes, both intentional and unintentional, that occur in the tape medium. This project aims to develop an automated process within the framework of the Moving Picture, Audio, and Data Coding by Artificial Intelligence (MPAI) Context-based Audio Enhancement (CAE) standard, which includes Audio Recording Preservation (ARP) as a use case. The methodology employs frame differencing techniques to compare consecutive video frames, enabling the detection of subtle alterations. Following this comparison, filtering and image processing operations are applied to highlight potential irregularities within the audio tape recordings. Various processed frames, including direct screenshots, difference frames and thresholded frames are extracted and a dataset is created to represent different types of irregularities. The content of the dataset is inputted into a ResNet-based deep learning module for automated classification of detected irregularities. The proposed method seeks to improve the automation and precision of irregularity detection in magnetic audio tape videos, advancing the field of digital archiving and preservation.

Acknowledgements

I would like to thank my supervisor, Professor Sergio Canazza, for his guidance, and my co-supervisors, Alessandro Russo and Matteo Spanio, for their invaluable support. Special thanks to my lab mate, Mehmet Öztürk, for his help throughout this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

From its invention by German Fritz Plaumer in 1928 until the advent of digital technologies in the 1980s, magnetic tape was one of the dominant media for audio recordings. Besides being used within professional studios for musical content, magnetic tape was also employed in diverse fields such as linguistic and anthropological research activities, which contributes to its significance as a part of cultural heritage [1]. Due to the deteriorating nature of the material, there is a risk of these culturally significant materials becoming obsolete. To mitigate this risk, ongoing digitization efforts are ongoing. The primary focus of these research efforts is the preservation of the content stored on these tapes and the preservation of their integrity.

Directly digitizing the audio content of every tape is an inappropriate method, as historically faithful preservation necessitates storing all information regarding the document, both audial and visual [2]. This includes the presence of both intentional and unintentional irregularities, such as annotations, chemical damage, and tape discontinuities, some of which can only be detected during the digitization process [3]. This introduces a challenge due to the vast amount of content and the necessity to automate this process. Addressing this challenge, Centro di Sonologia Computazionale (CSC) [4] has established the standard called MPAI-CAE ARP under the MPAI framework. This standard involves the active preservation of sound recordings, digitization, and ensuring long-term access. It also utilizes artificial intelligence to automate the detection and identification of irregularities. This thesis proposes an enhancement to the current methodology within this framework to improve the automated detection of magnetic tape irregularities.

1.1 Audio Preservation and Digitization

The degradation of analog carriers such as magnetic audio tapes can be slowed down but not stopped. To preserve the audio content, correct preservation policies are essential. The survival of the information contained in these carriers is possible only by transferring the information to new carriers [5]. Many organizations, such as the International Federation of Library Associations and Institutions (IFLA) and the International Association of Sound and Audiovisual Archives (IASA), have established standards for preservation specifications. According to IFLA, the preferred method of preserving the content of analog carriers is the digitization of the content [6]. However, digitization involves many considerations. IASA TC-03 [7] provides a standard for audio document preservation, specifying numerous aspects such as the digitization process, quality assessment, metadata documentation, storage, and accessibility.

In the context of audio preservation, metadata documentation is important. According to IASA TC-04 [8], metadata refers to structured information that describes and provides context for digital audio objects. This information is useful for their management, and long-term preservation. Examples of metadata include descriptive elements like title, creator, and keywords; technical details such as file format, sample rate, and duration; administrative information including preservation actions and rights management; and provenance data documenting the history and origin of the audio resource. Effective metadata documentation ensures accurate identification, accessibility and management of digital audio documents in preservation efforts.

1.2 Centro di Sonologia Computazionale

The Centro di Sonologia Computazionale (CSC) at the University of Padova is a prominent research institution specializing in music and computing. In the last few decades, CSC has focused on various aspects of music technology, including the preservation and restoration of historical audio documents, specifically the documents on analog magnetic tapes [9].

The inherent susceptibility to deterioration of magnetic tapes makes the preservation efforts urgent and crucial. The CSC's approach combines various fields such as musicology, philology, and information engineering to ensure accurate and comprehensive preservation of these culturally and historically valuable audio documents. The methodologies developed by CSC are applied in international projects and involve not only digitization of the audio content, but also metadata extraction and storage to maintain the integrity of both the audio content and its contextual information [2].

The preservation of magnetic tape documents at CSC involves a detailed approach to ensure the preservation and integrity of audio recordings. The main challenge in preserving analog magnetic tapes lies in capturing not only the audio signal but also the context and metadata of each tape. This includes the physical and chemical condition of the tape, the recording environment, and any alterations made on the carrier. CSC has developed a system that digitizes these tapes while preserving this additional information, consequently creating a comprehensive digital archive that reflects the original recordings as closely as possible. By doing this, the original intent and nuances of the recordings are kept intact. This process involves using advanced software tools to analyze and correct any errors introduced during digitization, such as speed variations and equalization inconsistencies [2].

1.3 MPAI-CAE ARP

A significant part of CSC's methodology is reflected in the MPAI's ARP (Audio Recording Preservation) use case. It was established to address the lack of a standard with software references in the audio document preservation. It focuses on digitizing audio documents and making sure that the preservation master file contains metadata and video of the tape in addition to audio in order to maintain the philological integrity of those documents [4]. Artificial intelligence modules are utilized for the analysis and correction of audio irregularities. This comprehensive approach enables the content, context, and historical significance of these recordings to be accurately preserved.

1.3.1 Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI)

Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) is an international non-profit organization that aims to develop standards for data coding using Artificial Intelligence (AI) [10]. These standards make it easier to transform data into formats that work well with a wide range of applications. The goal of MPAI is to develop a workflow of interchangeable and upgradeable AI Modules (AIMs) without changing the applications' fundamental logic [11]. This approach provides flexibility and it ensures the ongoing advancement of AI technologies.

The MPAI-AIF (AI Framework) is a standard designed to enable the creation and automation of mixed Machine Learning (ML), AI, and legacy data processing modules [11]. The first version of this standard has been adopted by IEEE as IEEE 3301-2022 [12]. Fig 1.1 illustrates the workflow of the MPAI-AIF framework [13].

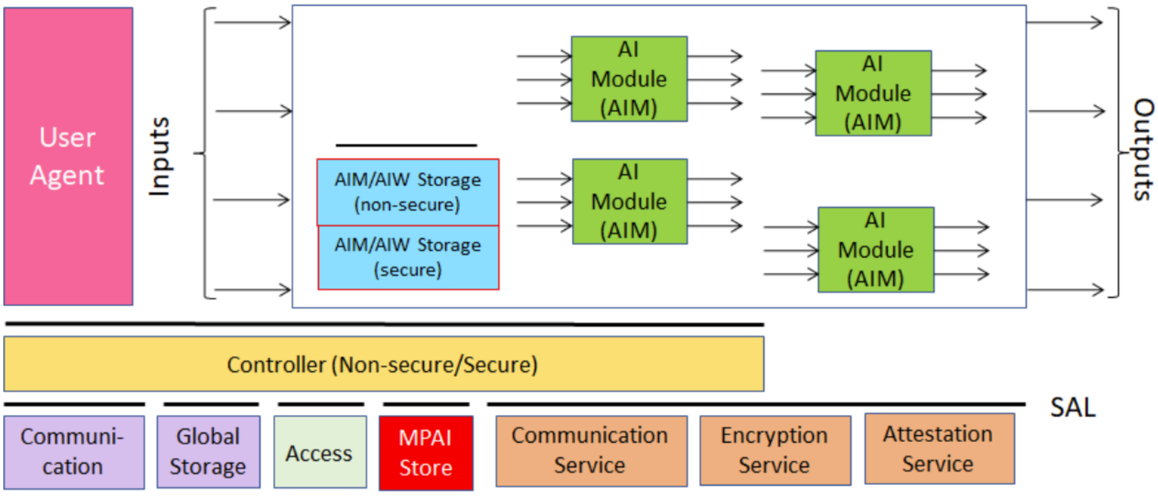


Figure 1.1: MPAI-AIF V2 Reference Model.

One of the key features of MPAI-AIF is interchangeability. AIMs can be replaced or upgraded without changing the overall logic of the application. This allows for continuous improvement on the application. Another goal of this framework is to provide compatibility and reusability by providing standardized interfaces for data exchange between AIMS. The framework is composed of several components, including management of AIMs, execution environment, storage, and access to data. This structure aims to support integration and operation of diverse AIMs [11].

1.3.2 Context-based Audio Enhancement(CAE)

Context-based Audio Enhancement (CAE) is a standard under MPAI and the first version of the standard has been adopted by IEEE as IEEE 3302-2022 [14]. MPAI-CAE standard aims to improve the user experience for a number of audio-related applications such as entertainment, communication, restoration etc. [15]. It includes four use cases designed to improve the quality of audio experiences in various contexts. These four use cases are described as follows:

- Emotion-Enhanced Speech (EES): This use case allows users to add emotional content to neutral speech, therefore enhancing the expressiveness of the speech. The users can transform plain speech into more expressive version by selecting an emotional tone. This use case can be useful in entertainment and communication where conveying emotions are crucial.

- Audio Recording Preservation (ARP): It focuses on preserving audio recordings, such as those on open-reel magnetic tapes. This use case involves creating digital versions that are suitable for long-term storage. It also ensures that these recordings can be played back correctly and ,in the cases where it is necessary, restored to maintain their quality and content.

- Speech Restoration System (SRS): It focuses on improving the clarity of a speech that has been degraded by noise or any other distortions. It enhances the intelligibility and quality of a distorted speech. This use case is particularly useful for applications like teleconferencing and archival restoration, where clear communication is essential.

- Enhanced Audioconference Experience (EAE): It aims to enhance the quality of audio in teleconferences by reducing background noise, echo, and other disturbances. This ensures that conversations are clear and effective, improving the overall experience and productivity of remote meetings.

1.3.3 Audio Recording Preservation (ARP) Use Case

Audio Recording Preservation (ARP) is one of the four use cases of MPAI-CAE standard. This use case focuses on preserving audio recordings stored on analog media, such as open-reel magnetic tapes. As mentioned earlier, open-reel tapes often contain crucial information, like splices, irregularities due to the physical and chemical characteristics of the tape and annotations made by composers or technicians. Properly documenting and preserving this information is essential for accurate playback and long-term preservation. Additionally, the ARP use case involves creating a digital copy for long-term preservation and an access copy that may require restoration to ensure correct playback.

AI Workflow (AIW) and Modules in ARP

The AI workflow for ARP consists of a structured sequence of operations performed by various AIMs. These AIMs work in unison to detect and classify audio and visual irregularities and restore the audio content. The process begins with digitizing the analog signal and capturing the video of the playback head. The Audio and Video Analysers then detect and classify any irregularities. The Tape Irregularity Classifier processes these classifications, and the Tape Audio Restoration module restores any corrupted audio. Finally, the Packager compiles all the relevant files into preservation and access copies.

In Fig 1.2, the AIW of ARP use case is shown.

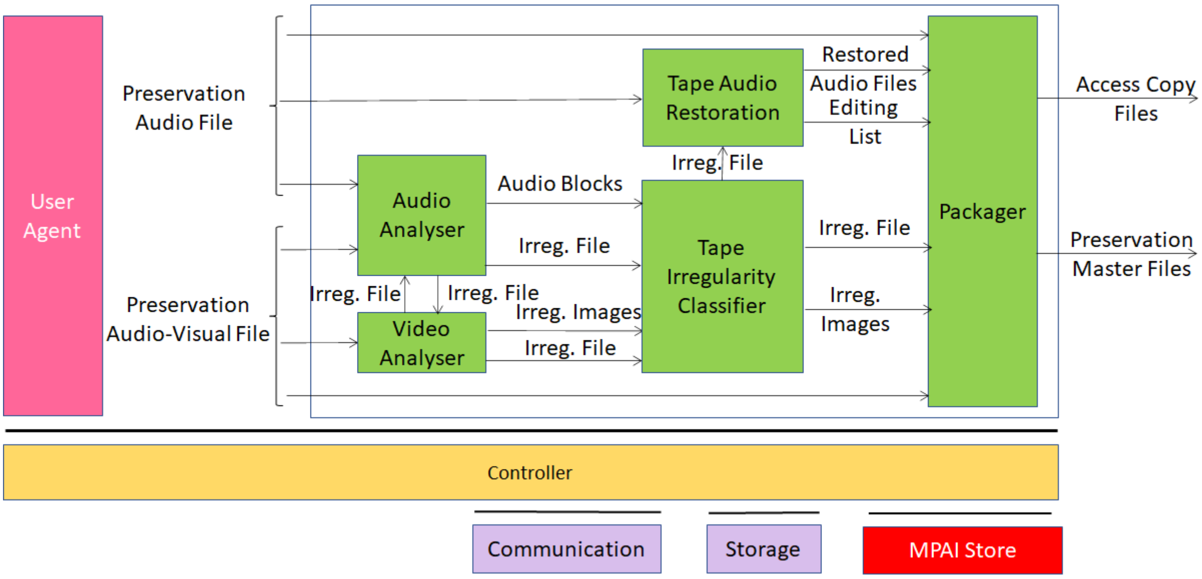


Figure 1.2: MPAI-CAE ARP Reference Model.

Audio Analyser

Audio Analyser extracts relevant audio fragments from the tape and classifies sections with low signal levels. It detects irregularities based on recording speeds and equalization curves, providing analyzed audio blocks and irregularity files to the Tape Irregularity Classifier. Audio Analyser works in unison with Video Analyser. It receives the irregularities file generated by Video Analyser and extracts audio files corresponding to the detected irregularities in the file. Finally, it combines the generated Irregularity File with the one that was obtained from the Video Analyzer and sends it to the Tape Irregularity Classifier along with the appropriate Audio Files.

Video Analyser

Video Analyser detects tape surface irregularities by utilizing computer vision algorithms. It sends the generated irregularity file to Audio Analyser and it receives the irregularities file generated by Audio Analyser. It extracts irregularity images corresponding to the detected irregularities in the file sent bu Audio Analyser. Similar to how Audio Analyser works, it combines the generated Irregularity File with the one that was obtained from the Audio Analyzer and sends it to the Tape Irregularity Classifier along with the appropriate irregularity images.

Tape Irregularity Classifier

Tape Irregularity Classifier processes the information provided by the Audio and Video Analysers, classifying irregularities for further processing and restoration. It ensures that all relevant irregularities are accurately documented and addressed. It sends the Irregularity File related to the selected Irregularities to Tape Audio Restoration module.

Tape Audio Restoration

Tape Audio Restoration module focuses on restoring audio segments that were corrupted or incomplete during digitization. It enhances the quality of the audio by detecting and correcting speed, equalisation and reading backwards errors in Preservation Audio File. Finally, it sends Restored Audio Files and Editing List to Packager.

Packager

Packager is responsible for producing Preservation Master Files and Access Copy Files. When assembling the Preservation and Access Copies, the Packager makes sure that all digital content, documentation, and metadata are appropriately arranged and stored. It creates comprehensive Preservation Master Files and Access Copy Files by combining files like the Preservation Audio File, Restored Audio Files, Editing List, Irregularity File, Irregularity Images, and the Preservation Audio-Visual File [4].

1.4 Automated Irregularity Detection

Automated irregularity detection is employed in audio preservation to identify and rectify various irregularities in audio carriers. These irregularities include splices, brands, chemical damage, biological contamination, etc. The irregularities on the tape can arise from various factors including physical damage to the recording medium, degradation over time and, in the context of tape music, intentional annotations by the composer. Detecting these irregularities is essential to ensure that the digitization process accurately captures the audio content without loss or distortion, thereby preserving the integrity and usability of historical and cultural audio archives.

Automation in irregularity detection is crucial due to the vast amount of audio data that needs to be processed. Manual detection and correction are time consuming and the process is prone to human error. Automated systems can process large volumes of audio data quickly and accurately. This efficiency is crucial in the context of cultural heritage, as it benefits institutions that manage extensive audio collections, such as libraries, archives, and research institutions .

The Video Analyzer is a software module developed by CSC for automated detection, designed in accordance with the guidelines of the MPAI-CAE ARP use case. It designed to detect significant frames from digital video footage, focusing primarily on the reading head of a tape recorder and the area under the pinch roller. The software uses frame differencing techniques, where significant frames are identified by comparing consecutive frames and detecting substantial color changes [3].

1.4.1 Irregularity Types

An irregularity, in the context of audio preservation, is defined as any deviation or anomaly on magnetic audio tapes that can affect the quality and integrity of the recorded sound [2]. Irregularities can be either intentional or unintentional.

Intentional irregularities include the alterations on the tape in the context of tape music. Tape music is a genre of electroacoustic music that involves the manipulation of recorded sounds on magnetic tape. Therefore, it includes modifications on the tape made by composers. These modifications include annotations and splices and they are integral to the creative process. Composers like Pierre Schaeffer and Karlheinz Stockhausen used splicing and looping techniques to create new musical forms and textures. Annotations made directly on the tape provide instructions for playback and synchronization. They serve as a guide for the performers and technicians [16]. Since these alterations carry significant information, their accurate preservation is culturally and historically crucial

Unintentional irregularities are deteriorations that develop over time. These irregularities include physical damage, such as scratches, splices, and tape breakage; chemical damage due to change in temperature and humidity; and contamination, such as dust and dirt accumulation on the tape surface that leads to playback issues [17].

The IASA Cataloguing Rules [18] provide an extensive list of conditions that can appear on a audio tape. Not every condition appears equally often in magnetic audio tapes. To make the detection and classification process easier and more feasible, a simplified approach is adopted. The simplified classification scheme used in [2] includes four primary classes:

- Splice: Connections where two segments of tape have been joined together with adhesive tape
- Shadow: Ghosting or imprints left on the tape
- Brand: Identification marks or logos printed on the tape by the manufacturer.
- Ends-of-Tape: The point where the tape is not under tension.

In the scope of this thesis, “ends-of-tape” class is removed as the ends-of-tape instances are detected by a different region of interest than the other irregularities (this will be explained in Section 2.2.1). The classes are considered for this research are “splice”, “brand” and “shadow”.

Chapter 2

Irregularity Detection

In accordance to the guideline of MPAI-CAE ARP use case, the Video Analyser module is responsible for detecting the irregularities on a tape [4]. The irregularity detection is done by the employment of various computer vision algorithms. This thesis proposes an improvement in the functioning of this module.

2.1 Existing Method

In the scope of MPAI-CAE ARP standard, the Video Analyser developed by CSC [19] employs frame by frame comparison between consecutive pairs of frames. After the comparison, the frames that display significant differences compared to their predecessors are captured as potential irregularities. The algorithm of the Video Analyser works as follows:

1. Identification of Regions of Interest

2. Analyzing the video sequence and comparing the consecutive pairs of frames by counting the numbers of dissimilar pixels.

3. If the count of non-identical pixels exceeds the predetermined threshold, irregularity is detected.

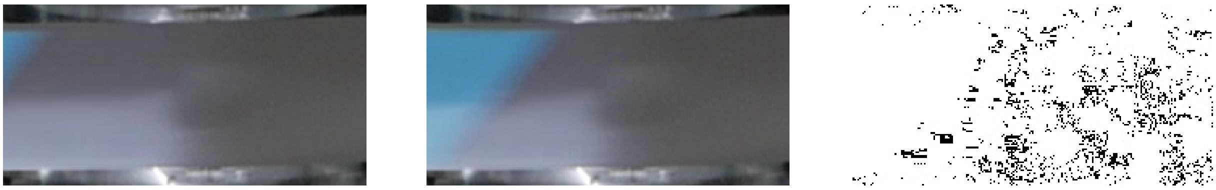
The regions of interest are identified by using Hough Transform and Speeded-Up Robust Features (SURF). This will be discussed further in Section 2.2.1 where its implementation and results will be examined in detail.

After the ROIs are found, the video is traversed and consecutive pairs of frames are compared pixel by pixel in the tape area. For every pair of frames, a binary comparison image is generated with non-identical pixels are represented by black pixels and identical pixels are represented by white pixels.

The generation process of comparison image is outlined as follows:

$$D(i,j) = \begin{cases} 255 & \text{if } C_{\text{red}}(i,j) - P_{\text{red}}(i,j) = 0 \text{ and} \\ & C_{\text{green}}(i,j) - P_{\text{green}}(i,j) = 0 \text{ and} \\ & C_{\text{blue}}(i,j) - P_{\text{blue}}(i,j) = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $i = 1, \ldots, n$ and n is the number of rows in the matrix, $j = 1, \ldots, m$ and m is the number of columns in the matrix. $C_{\text{red}}, C_{\text{green}}$, and C_{blue} are the current frame matrices for the red, green, and blue color channels respectively, and $P_{\text{red}}, P_{\text{green}}$, and P_{blue} are the previous frame matrices for the red, green, and blue color channels respectively.



(a) Previous frame (b) Current frame (c) Comparison image

Figure 2.1: The output of existing algorithm for two consecutive frames

It has been observed that approximately 80% of the pixels consistently show variations and 20% of the pixels match with the pixels on the previous frame. In the case of an irregularity, the number black pixels increase significantly. When the quantity of black pixels exceed a predetermined threshold, the irregularity is detected.

Despite its utility, this method has limitations. The generation process of the comparison image analyzes any change on RGB channels however magnitude of the difference is disregarded. Since the video footage of an audio tape displays a constant movement on the tape area, the small movements and vibrations on the tape area are registered as non-identical pixels. This rises the likelihood of errors. In addition, not distinguishing between significant differences and negligible ones increases the susceptibility to errors under gradual change of lighting.

In response to these limitations, this thesis proposes an alternative solution for a more robust irregularity detection. The methodology involves employing traditional frame differencing coupled with a more versatile approach to thresholding.

2.2 Methodology Overview

The proposed method incorporates the framework of the existing method where consecutive pairs of frames are compared while traversing the video. The novelty underlying this method is its approach to the presented problem as a motion detection problem. Despite the existence of constant movement in the entire ROI throughout the video, a tape region with no irregularity displays a static image-like appearance in motion. This facilitates the execution of a traditional motion detection algorithm with frame differencing wherein the irregularities are considered as moving objects and regular region is considered as background.

Frame difference technique adapted by [20] consists of two main stages. The first stage is generating a difference image by calculating the absolute difference between two consecutive frames. The second stage is thresholding the difference image to create a meaningful binary image for determining whether the given frame displays an irregularity or not. To separate the background and the regularity properly, a thresholding technique implemented. This technique consists of Otsu's method and global thresholding coupled with opening operation as a post processing step.

The implemented method works as follows:

1. Identification of Regions of Interest(Adapted from the state of art methodology)

2. Absolute difference of pairs of consecutive frames are calculated to obtain a difference image.

3. Based on the standard deviation of the difference image, it is decided if the frame is to be evaluated or not.

4. Otsu's method is used to compute the threshold. Based on the obtained value, either global or Otsu's threshold is picked for thresholding.

5. Thresholding is applied on the difference image to obtain a binary Motion image.

6. Opening operation is applied on the Motion Image to get a better view on the potential irregularity and clean the image from the spurious detections caused by vibration of tape and other external factors.

7. If the number of white pixels exceeds the given threshold, the frame is detected as an irregularity.

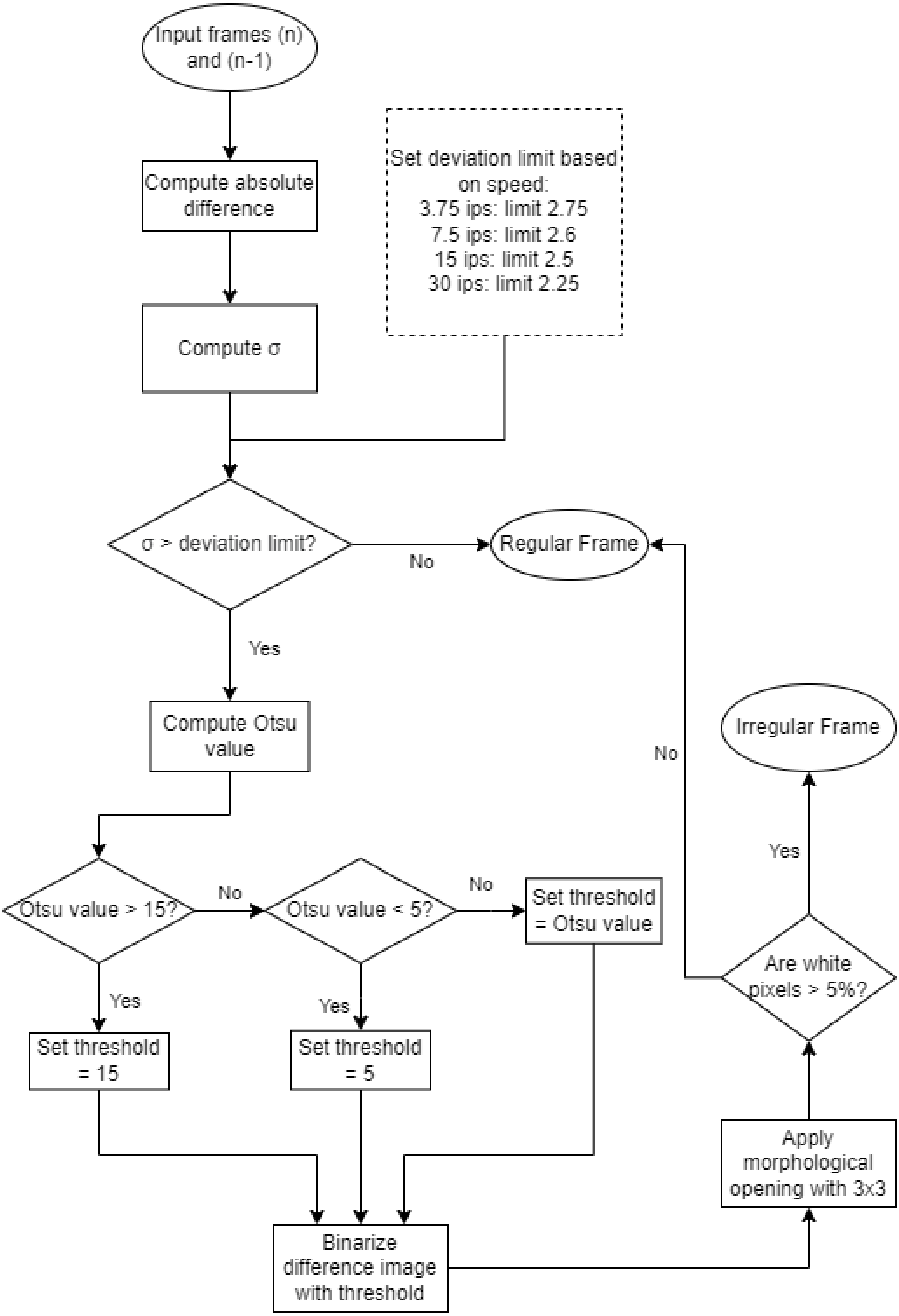


Figure 2.2: Flowchart of the new frame differencing method.

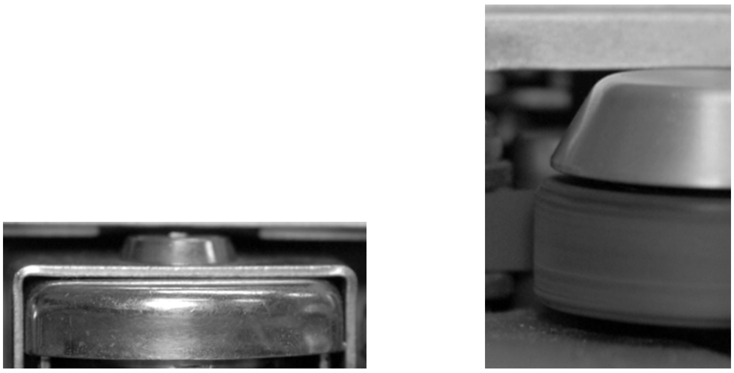
The threshold values used for the tests are presented under the respective sections.

However, it should be noted that, with the growing data, the numbers are susceptible to change in the future iterations.

2.2.1 Region of Interest Detection

There are two ROIs to be detected, tape area and capstan area. While the capstan area is used only for early stop in the software at the end of the tape, the tape area is where the irregularities are detected. These areas are found with the aid of two template images belonging to reading head of tape machine and the capstan. Tape area is detected by using Generalized Hough Transform(GHT). GHT permits the detection and matching of a specific pattern on the image [21]. With its utilization on the template image, the specific pattern on the image is detected. Following this, the potential matches across the image are found and the most probable location is chosen. After the reading head is found, the tape area is decided based on the proportions of the reading head and predetermined coordinates.

Capstan area, in contrast, is detected by using Speeded-Up Robust Features(SURF). SURF is a feature detection and description algorithm and it detects the keypoints and descriptors on both the main image and template image [22]. Descriptors of keypoints extracted from the template image are compared with the descriptors extracted from the main image. To address the risk of obtaining an inaccurate ROI when the number of matches is not enough to signify detection, an additional threshold of 25 is added to stop the process.



(a) Template image of reading head

(b) Template image of capstan

Figure 2.3: Gray-scaled template images.

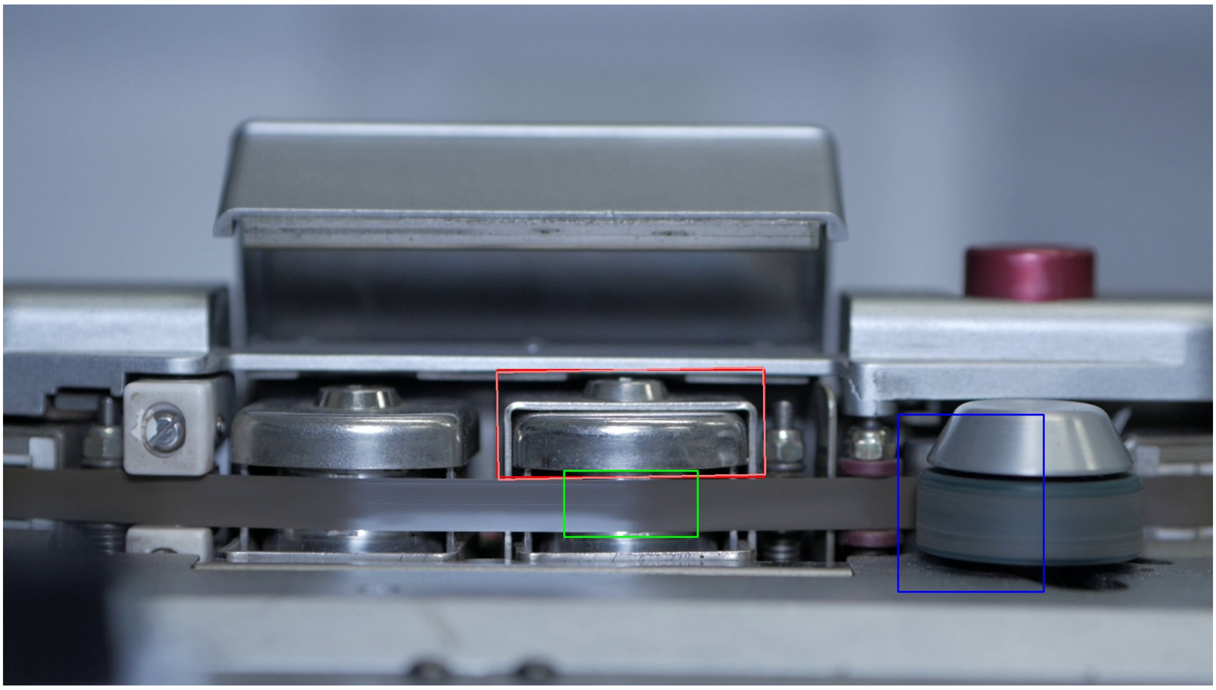


Figure 2.4: Tape areas shown on the main image

2.2.2 Absolute Difference

In the context of frame differencing and motion detection, the color variations may not consistently be an indicative of motion. Instead, the variations of intensity play a more crucial role for detection of the motion. To focus on the intensity difference between the frames, the images are converted to gray-scale.

The absolute difference is the operation where a difference image is generated based on the discrepancies between two input images [20]. The difference image contains pixels with intensity values representing the magnitude of the difference of intensity between the corresponding pixels in the input images. As a result, in the difference image, the pixels with small variations have darker pixels as they are close to 0 and pixels with significant variations have lighter pixels.

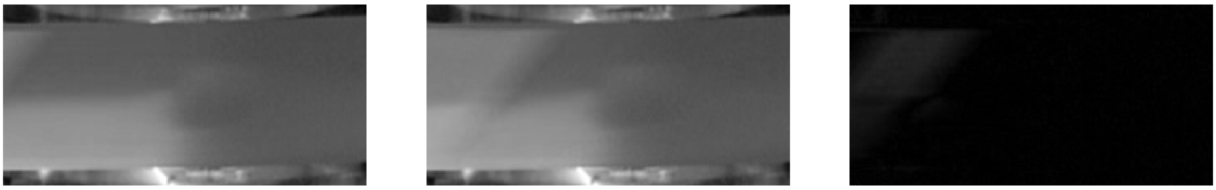
Let $I(x, y)$ be the intensity of a pixel at position (x, y) in the image. Suppose the image has dimensions $W \times H$, where W is the width and H is the height. The conversion to gray-scale of a colored image is defined by OpenCV [23] is as follows:

$$I(x, y) = 0.2989 \cdot R(x, y) + 0.5870 \cdot G(x, y) + 0.1140 \cdot B(x, y)$$

The absolute difference at pixel (x, y) between consecutive frames is defined as the absolute value of the difference between the pixel intensities of the current frame I_n and the previous frame I_{n-1}:

$$\Delta(x, y) = |I_n(x, y) - I_{n-1}(x, y)|$$

In the Fig 2.5, the gray scaled versions of the images 2.1a and 2.1b and the generated difference image are shown.



(a) Gray-scaled previous frame

(b) Gray-scaled current frame

(c) Difference image

Figure 2.5: The absolute difference of two consecutive frames

2.2.3 Selection of Frames to be Evaluated

In order to separate the irregularity and regular tape area, Otsu's threshold is utilized. However, the computation of Otsu's threshold is an iterative process where each color value on the histogram of the image is traversed [24]. If employed on every frame, this process would cause computational burden on the program. To prevent this, an additional decision-making step based on standard deviation is employed before the computation of threshold.

Standard deviation is computed for each difference frame while the video is being traversed. Only in the case that standard deviation is larger than a set threshold, the process continues. Otherwise, the frame is taken as a regular frame.

The rationale for selecting standard deviation as a decision-making step is rooted in its value as a metric for assessing the distribution of pixel intensities. Greater standard deviation indicates a larger amount of variation of pixel intensities, which in turn signifies a greater difference between two consecutive frames. Moreover, unlike Otsu's method, this process is not iterative and consequently, it does not impose a significant computational burden on the program.

The mean intensity μ of the image can be calculated as:

$$\mu = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \Delta(x, y)$$

The standard deviation σ of the image can be calculated as:

$$\sigma = \sqrt{\frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} (\Delta(x,y) - \mu)^2}$$

After the tests with various data, it is observed that the average standard deviation tend to get smaller as the tape speed increases. The reason behind this decrease is the increase of blur on the images. Absolute difference between images with motion blur yields smaller standard deviation as the pixels that represent different areas have smaller intensities. This is further discussed in Section 2.5. In response to this decrease, different standard deviation thresholds are employed. For the tests cases, the used thresholds are: 2.25 for 30 ips, 2.5 for 15 ips, 2.6 for 7.5 ips, 2.75 for 3.75 ips.

2.2.4 Threshold Computation

For the difference images that fall outside the range defined by the standard deviation threshold, a global threshold value T is assigned. First, Otsu's method is employed and a threshold value is obtained based on the histogram of the image. If the obtained threshold value is between the desired range, T is assigned this value and is applied in the following step. If the value is out of the desired range, it is assigned a predetermined integer. T is assigned the upper bound in the cases where the value exceeds the upper bound and is assigned the lower bound in the cases where it is under the lower bound. The rationale for using Otsu's method to find a suitable threshold value and the necessity of the upper and lower bound will be discussed further in Section 2.3.

Let T_o denote a threshold value obtained by the Otsu's method, U denote the upper bound and L denote the lower bound. The threshold T is determined by the following conditions:

$$T = \begin{cases} U & \text{if } T_o > U \\ T_o & \text{if } L \leq T_o \leq U \\ L & \text{if } T_o < L \end{cases}$$

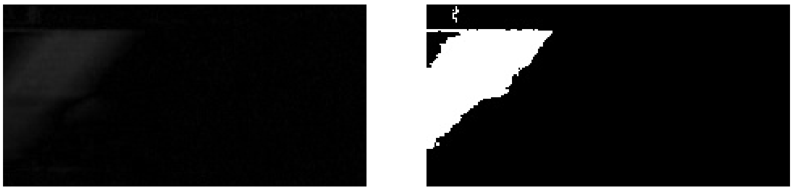
It has been observed that assigning 15 to the upper bound and 5 to the lower bound yields positive results.

2.2.5 Thresholding the Image

Thresholding is a process to generate a binary image based on a threshold value T. Based on the threshold value obtained in the previous step, a thresholded binary image called motion image is produced. Each pixel that is greater than or equal to T is assigned an intensity value of 255 which corresponds to a white pixel. Each pixel under T is assigned an intensity value of 0 which corresponds to a black pixel. Thresholding operation based on the threshold value T can be described as:

$$M(x,y) = \begin{cases} 255 & \text{if } \Delta(x,y) \geq T \\ 0 & \text{otherwise} \end{cases}$$

In Fig 2.6, the resulting motion image is shown. The thresholding was done with the value obtained with Otsu's method.



(a) Difference Image

(b) Motion Image

Figure 2.6: The difference image is binarized with T = 13

2.2.6 Opening the Image

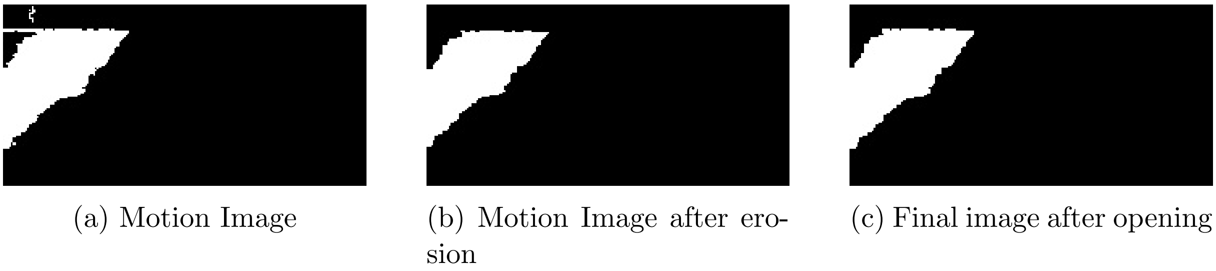
After obtaining the motion image, it is possible to observe the existence of some white pixels that are not a part of the main irregularity. These unrelated motion artifacts are likely to be caused by the vibration of the system or some external lighting change. In order to give the irregularity a clearer shape and mitigate the impact of these white pixels on the evaluation phase, opening operation is applied.

This operation is composed of two steps. First step is erosion to eliminate the unrelated artifacts. The second step is dilation to recover the size and shape of the main irregularity.

Let M be the binary motion image, O be the binary final image, and let S be the structuring element for the opening operation. The opening of M by S is denoted as $M \circ S$ and defined as:

$$O(x,y) = (M \circ S)(x,y) = (M \ominus S) \oplus S$$

For the video footage with 1920x1080 and 720x576 resolutions under the disposal of CSC, the structuring element used for opening operation is a 3x3 kernel.



(a) Motion Image

(b) Motion Image after erosion

(c) Final image after opening

Figure 2.7: Input image and two steps of opening operation with a 3x3 kernel

In Fig 2.7, results of the two steps are shown. 2.7a is the motion image obtained in the previous step. 2.7b is the image obtained after the erosion and 2.7c is the final image after the dilation.

2.2.7 Evaluation of the Frame

The final step consists of simply counting the number of white pixels and comparing it to the threshold percentage set by the user. If the number of white pixels exceeds the set threshold, irregularity is detected. If not, the frame is considered a regular frame.

The number of white pixels on the opened image can be described as follows:

$$\text{White pixel count} = \sum_{x=1}^{W}\sum_{y=1}^{H}\begin{cases}1 & \text{if } O(x,y) = 255\\ 0 & \text{otherwise}\end{cases}$$

Let threshold be the desired percentage threshold. The white pixel threshold can be calculated as:

$$\text{Threshold} = \frac{\text{thresholdPercentage} \times H \times W}{100}$$

It has been observed that the threshold of 5% yields good results and effectively captures the targeted irregularities. The details of the selection of white pixel threshold is discussed further in Section 2.4.

2.3 Selection of Threshold

For accurate detection of irregularities, it is crucial to implement an appropriate thresholding strategy. To achieve this objective, a technique that integrates Otsu's method and global thresholding is employed.

2.3.1 Otsu's Method

Introduced by Nobuyuki Otsu [24] in 1979, this method computes the intensity threshold that separates the pixels in the given image into foreground and background. This approach is beneficial in the context of motion and change detection because foreground is considered the moving object or the irregularity and the background is considered the static background or regular area.

The threshold is decided based on an iterative process where the histogram of the image is traversed from 0 upwards. In every iteration, the pixels are divided into two classes based on the value of the given iteration and the intra-class variance is computed. The intensity value that maximizes the inter-class variance and consequently minimizes the intra-class variance is returned as the threshold value. The pixels in the image are divided into two groups. The pixels that exceed the threshold belong to foreground, the pixels under the threshold belong to background.

Let T be the threshold value, $\omega_1(T)$ and $\omega_2(T)$ b the probabilities of the two classes separated by threshold T and $\mu_1(T)$ and $\mu_2(T)$ are the means of the two classes. Otsu's method computes the threshold T by maximizing the inter-class variance σ_b^2. Simplified version by [25] given by:

$$\sigma_b^2(T) = \omega_1(T) \cdot \omega_2(T) \cdot [\mu_1(T) - \mu_2(T)]^2$$

The threshold T_o is the one that maximizes $\sigma_b^2(T)$:

$$T_\mathrm{o} = \arg\max_{T} \sigma_b^2(T)$$

In Fig 2.8, for the difference image in Fig 2.6 the threshold value obtained with Otsu's method is shown on the histogram.

2.3.2 Upper Bound for Otsu's Threshold

In the cases where difference images have high intensities on certain areas, Otsu's threshold increases beyond the desired level. As a result, only the areas with high intensities are taken as white pixels and other parts of the irregularity are filtered out. To address this problem, an arbitrary upper bound is assigned and in the cases where Otsu's threshold

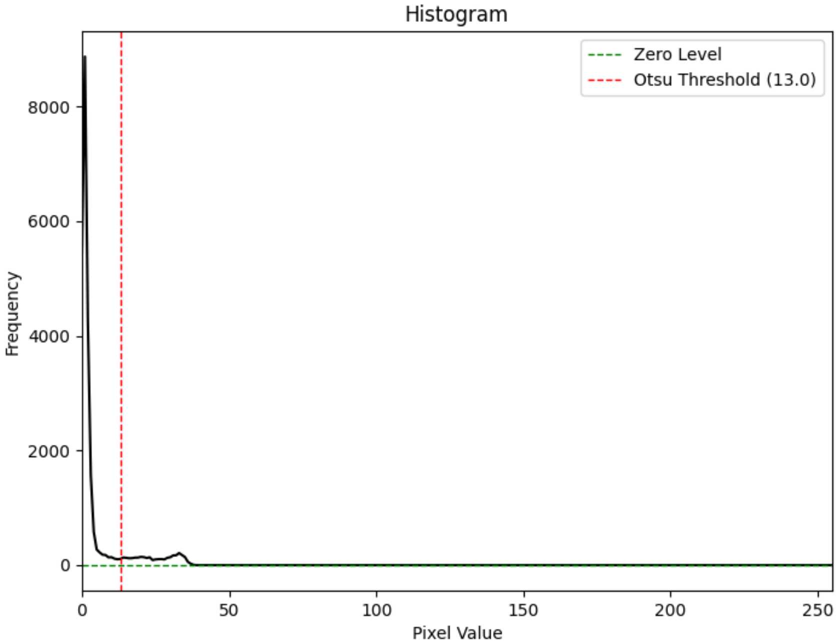
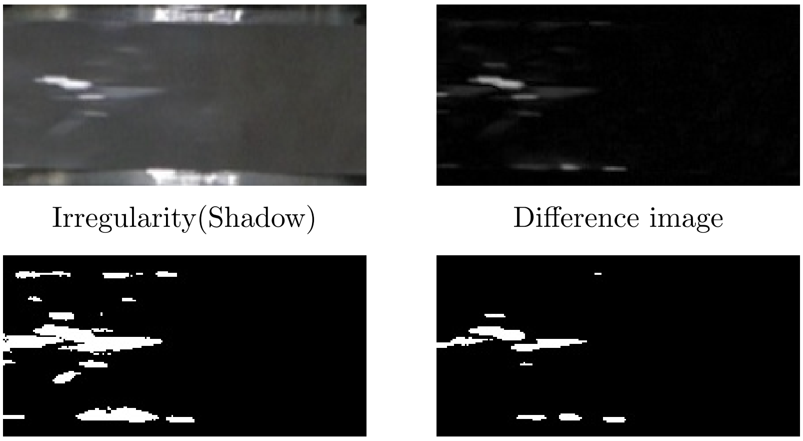


Figure 2.8: Histogram of the image 2.6.

exceeds this value, the binarization is performed with this upper bound.

As mentioned earlier, 15 as the upper bound yields the best results. In Fig 2.9, it is shown that some small areas of the difference image has intensity value significantly higher than other areas. When Otsu's method is applied, the obtained threshold is 35. However, it can be seen that this number is higher than desired as it eliminates many pixels that should be in the motion image. As a result, the pixel number cannot pass the threshold if Otsu's method is applied.



Irregularity(Shadow)

Difference image

Thresholded with T=15, White Pixels: 6.75%

Thresholded with T=35, White Pixels: 2.68%

Figure 2.9: Comparison of thresholding with T=15 and Otsu's threshold T=35

On the Histogram in Fig 2.10, the number of pixels that would be omitted if Otsu's method was used can be seen.

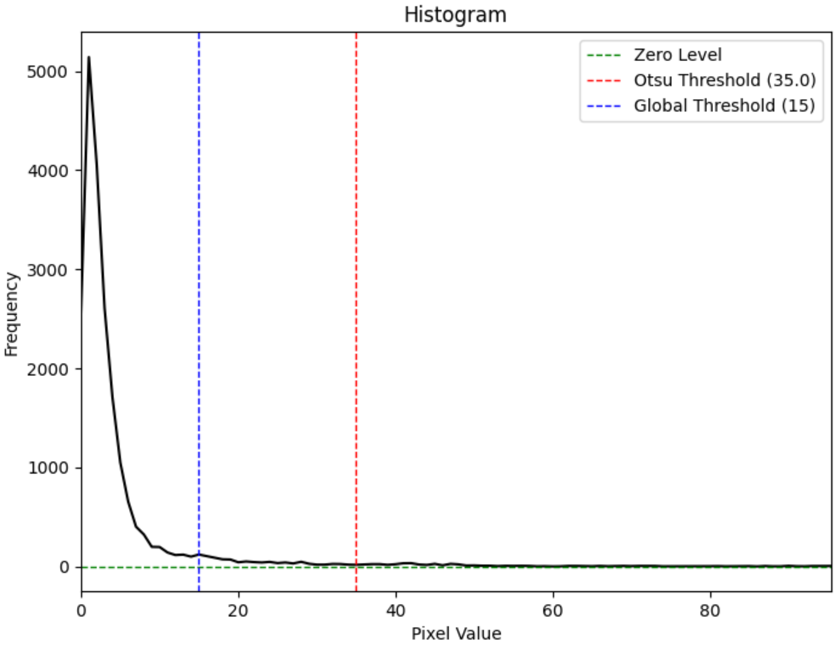


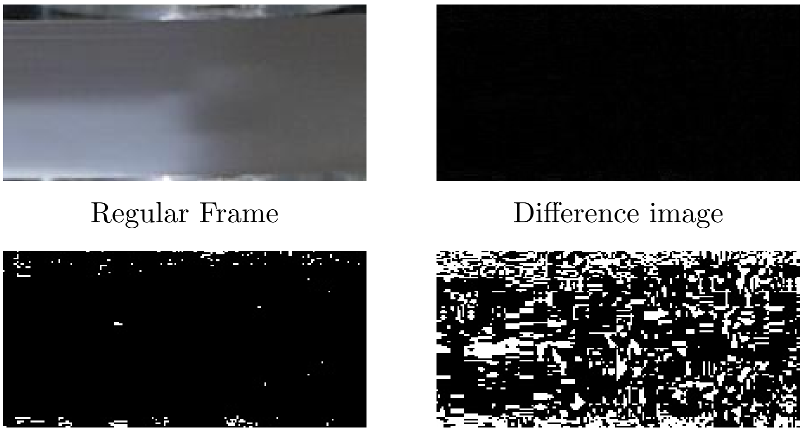
Figure 2.10: Histogram of the image in Fig 2.9 (zoomed in)

2.3.3 Lower Bound for Otsu's Threshold

Without the presence of an irregularity in the image, the difference image has low intensity across all its regions. In this scenerio, Otsu's method obtains a threshold value significantly lower than the desired level. When this threshold is used for binarization, many pixels with low intensities will assigned white pixels. This may result in false positive. Similar to the upper bound, an arbitrary lower bound is assigned to ensure the threshold value remains pertinent.

It has been observed that assigning the lower bound to 5 yields good results that eliminates some false positives and protects the integrity of image. However, it is important to note that the criterion of standard deviation mentioned in Section 2.2.3 diminishes the necessity of this lower bound as many regular frames will be eliminated due to their distribution of pixel intensities and low standard deviations.

In Fig 2.11, the results show the outcome when Otsu's method is applied to a regular frame. The obtained threshold is only 1. When this threshold is used for binarization, the generated motion image loses the integrity and generated a motion image filled with artifacts without any significance. When the lower bound is used, on the other hand, the number of white pixels are much lower than an the limit necessary for irregularity detection.



Regular Frame

Difference image

Thresholded with T=5, White Pixels:1.29%

Thresholded with T=1, White Pixels: 28.45%

Figure 2.11: Comparison of thresholding with T=5 and Otsu's threshold T=1

In Fig 2.12, the binarization of the same image can be observed on the histogram.

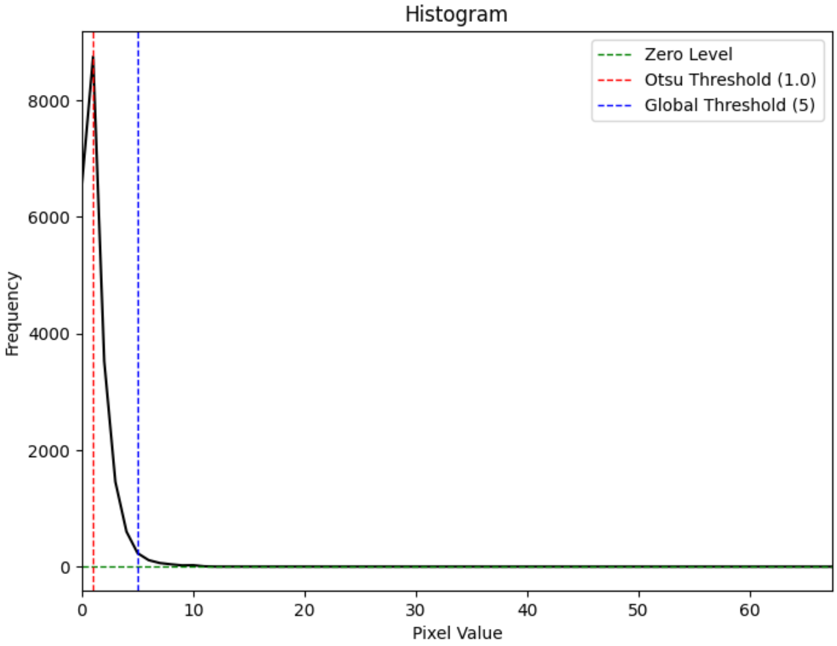


Figure 2.12: Histogram of the difference image in Fig 2.11 (zoomed in).

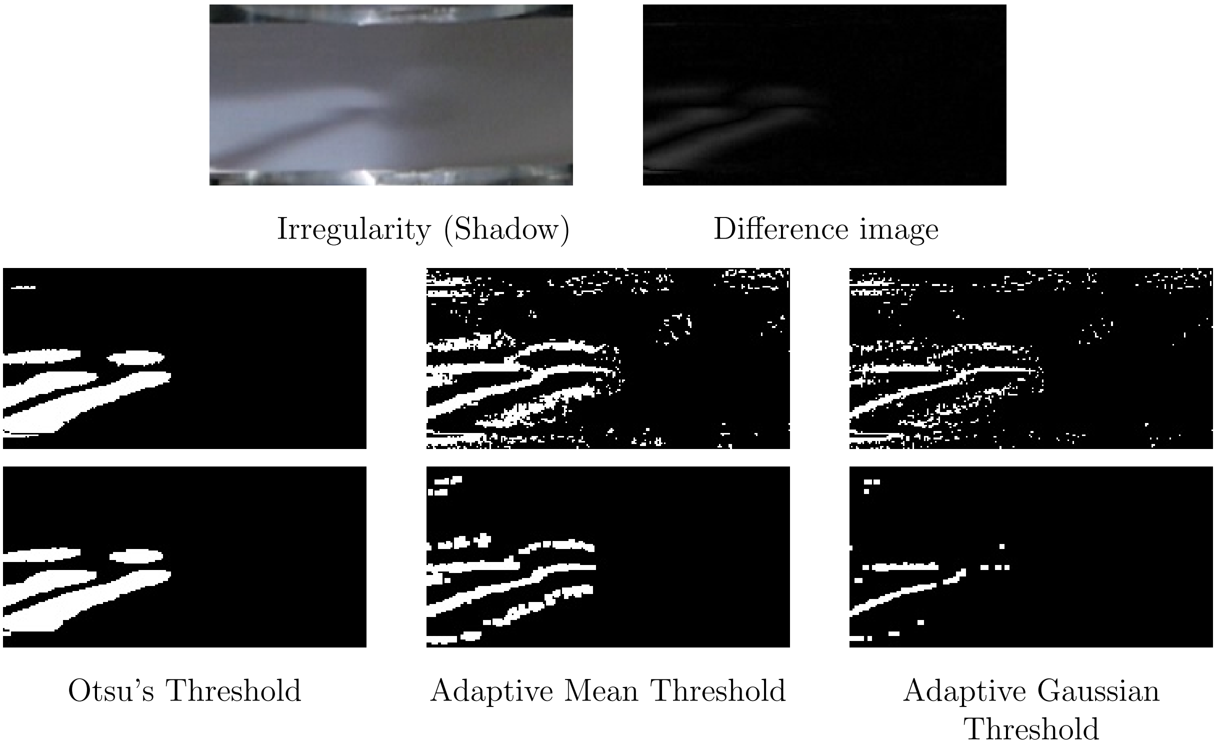
2.3.4 Comparison with Adaptive Thresholding

Adaptive thresholding is a technique that is widely used in image processing. This technique binarizes the images locally and allows for managing of the images under an unbalanced distribution of luminance.

The two main approaches of this technique are adaptive mean thresholding and adaptive Gaussian thresholding. Adaptive mean thresholding works by computing the local mean intensity around the pixel. Adaptive Gaussian thresholding, on the other hand, works by computing a weighted sum of pixel intensities around the pixel by using a Gaussian window.

Both approaches were considered to be employed in the proposed solution. However, despite their utility in image binarization, these approaches failed to address one of the determining criteria of the method. After the binarization with adaptive thresholding, the information regarding the area of the irregularity is lost. This made it necessary to implement a technique that is global in nature in terms of how it is applied to the given image.

In Fig 2.13, it can be seen that the shape and the size of the irregularity are not preserved when adaptive thresholds are applied. In addition, the motion artifacts are introduced.



Irregularity (Shadow)

Difference image

Otsu's Threshold

Adaptive Mean Threshold

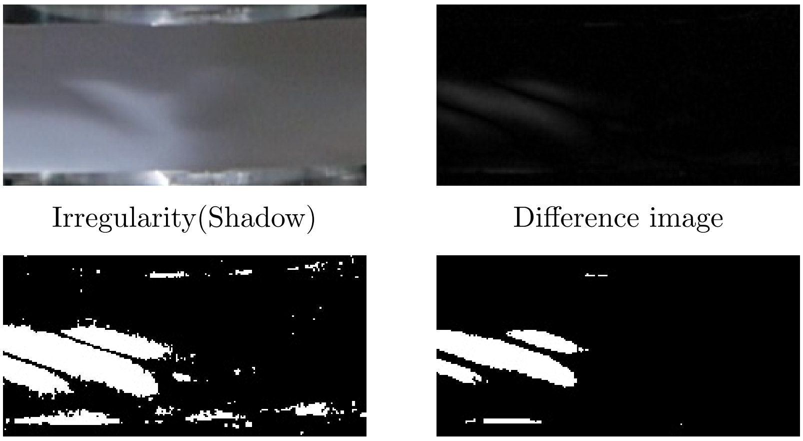
Adaptive Gaussian Threshold

Figure 2.13: Difference image of a shadow binarized with different techniques followed by opening

2.3.5 Comparison with Global Thresholding

As discussed earlier, Otsu's method is inherently global. However its approach is adaptive in how the threshold value is calculated based on the histogram. It was considered to remove this adaptive quality and to use global thresholding with a predetermined integer. However, due to varying nature of the irregularities and differing tape speed, no integer has been able to effectively capture the targeted irregularities. Therefore, it was deemed necessary to implement a technique with an adaptive nature such as Otsu's method.

In Fig 2.14, it can be seen that the motion image generated by Otsu's threshold preserves the integrity of the irregularity better as global threshold captures many motion artifacts that are not intended. In Fig 2.15, the global threshold is increased to prevent this outcome. But in this case, smaller irregularities are not captured as they do not exceed the white pixel threshold.



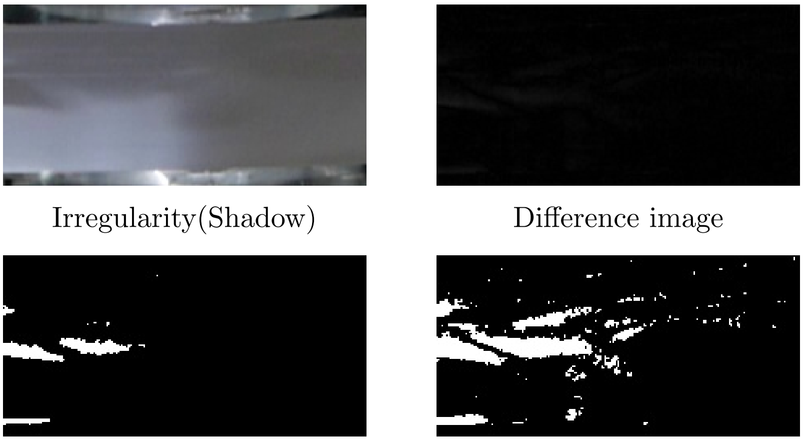
Irregularity(Shadow)

Difference image

Global threshold, T=5

Otsu, T=13

Figure 2.14: Comparison of thresholds of global T=5 and Otsu's method



Irregularity(Shadow)

Difference image

Global threshold, T=10

Otsu, T=5

Figure 2.15: Comparison of thresholds of global T=10 and Otsu's method

2.4 White Pixel Threshold Percentage

During the tests, it has been observed that 5% is an appropriate threshold to detect irregularity frames based on white pixel percentages. Several analytical methods such as histogram analyses, cumulative distribution function and sensitivity analysis have been employed to further verify the effectiveness of this threshold value. Fig 2.16, Fig 2.17 and Fig 2.18 show the application of these analytical methods on a typical video with the playback speed of 7.5 ips and features irregularities from splice and shadow classes.

During the histogram analyses, it is observed that the majority of frames percentages significantly below the 5% threshold. This distribution suggests a natural separation around the 5% mark. In Fig 2.16, this separation can be seen. This separation may occur at lower values for the videos that feature irregularities with small intensities. In order to avoid possible false positives, the possible highest separation is taken into account.

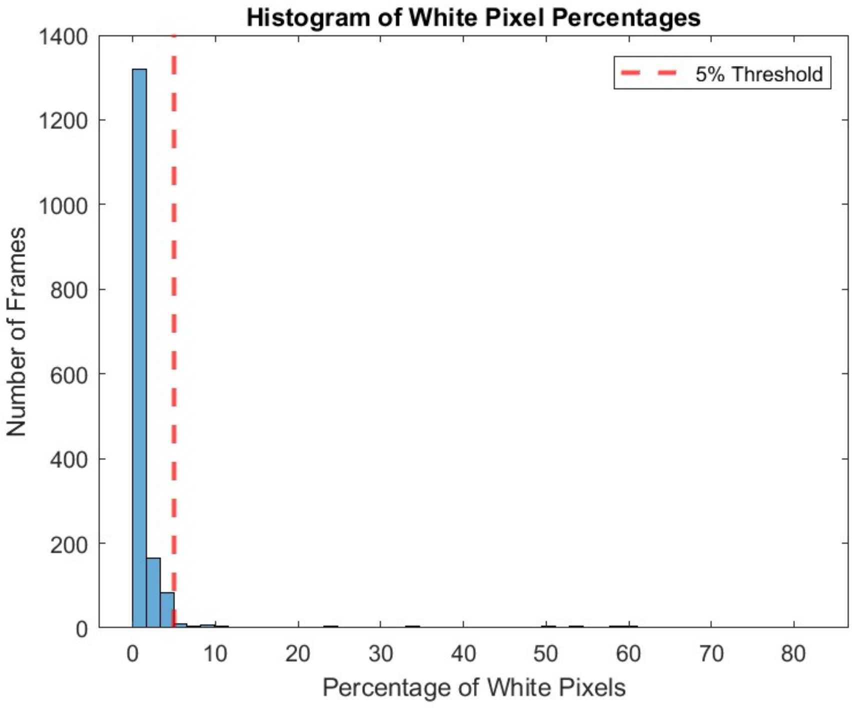


Figure 2.16: Histogram of white pixels of a sample video with 7.5 ips.

In Fig 2.17, The cumulative distribution function (CDF) plot reveals that a large number of frames fall below the 5% mark. This steep rise before the threshold means that most frames are considered regular, while those above this mark are likely irregular.

Finally, the sensitivity analysis shows how changing the threshold affects the number of irregular frames detected. Fig 2.18 shows that as the threshold decreases, the number of detected irregular frames increases sharply. This indicates that the 5% threshold is at

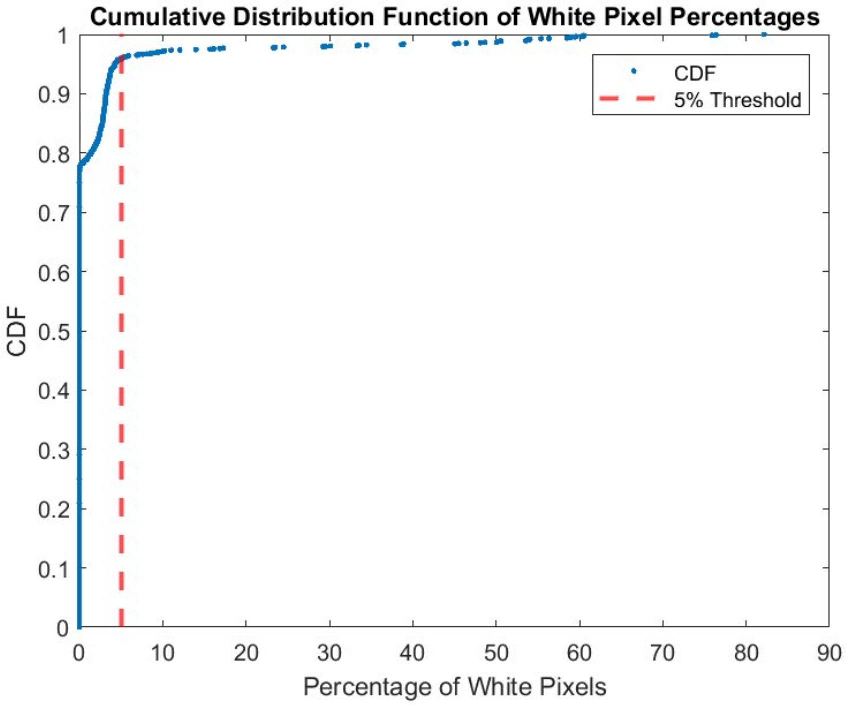


Figure 2.17: CDF of white pixel percentages a sample video with 7.5 ips.

a point where it effectively captures irregular frames without including too many regular ones.

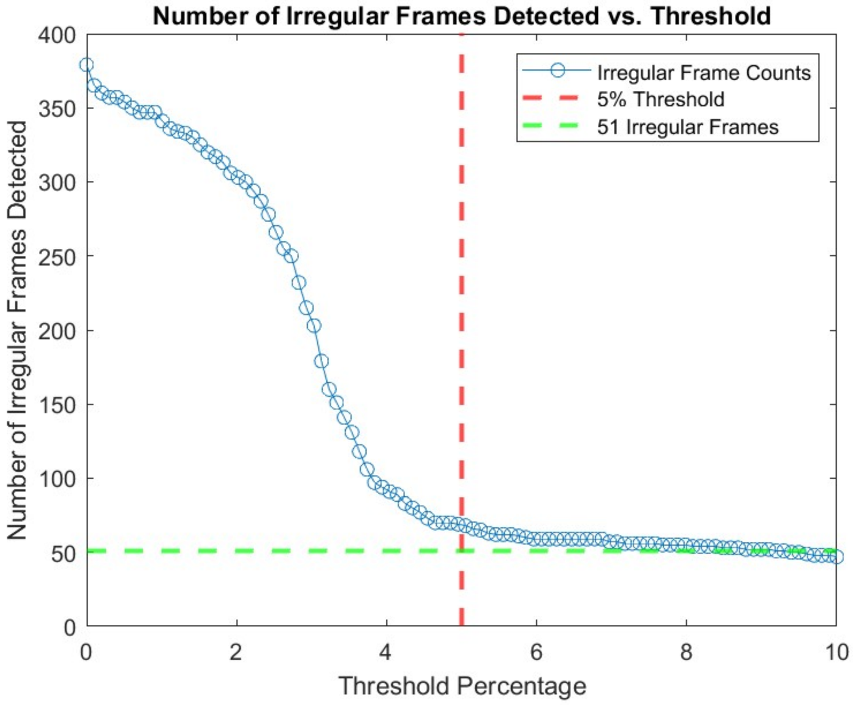


Figure 2.18: Sensitivity analysis of a sample video with 7.5 ips. The real number of irregular frames is 51.

2.5 Effect of Tape Speed on Standard Deviation

Since each tape has its own designated playback speed, observing the effects of speed in a controlled environment is challenging. This test was conducted to address this issue. Four different tapes, all produced by the same manufacturer, are used to ensure consistency in the ”brand” irregularities. The videos were recorded under similar lighting conditions to control for environmental factors. Notably, the tapes do not include irregularities classified under ”shadow” or ”splice” categories, allowing for a clearer analysis of how speed affects pixel intensity variations.

The analysis of standard deviation across varying video speeds reveals a relationship between speed and intensity variability. Table 2.1 shows the means and standard deviations of the standard deviations of pixel intensities. The mean of the standard deviations does not provide clear insights due to its dependency on the presence of irregularities. However, the standard deviation of the standard deviations is generally lower at higher speeds and this suggests that the motion blur introduced by higher speed options tends to smooth out intensity variations and reduce the fluctuation in pixel intensities. Alternatively, lower speed options show higher variability in their standard deviations, reflecting greater fluctuations in pixel intensity.

Speed (ips)	**Mean of Std Dev**	**Std Dev of Std Dev**
3.75	1.8409	1.7108
7.5	1.5730	1.1842
15	1.4296	0.7123
30	1.3997	0.4862

Table 2.1: Mean and standard deviation of the standard deviations of pixel intensities across different playback speeds.

Fig 2.19 shows that the overall distribution of standard deviations is narrower at higher speeds and the presence of outliers is more pronounced at lower speeds. This observation further confirms the impact of motion blur. It indicates that at lower speeds, the standard deviation is more sensitive to sudden intensity changes and this results in higher values for outliers. Alternatively, at higher speed, the effect of motion blur appears to mitigate these irregularities and reduce the magnitude of outlier values. The median value also increases with speed, likely due to the motion blur. As the speed increases, the increased motion blur averages out intensity variations and shifts the median higher. At the lowest speed, the box plot displays numerous outliers with higher values, indicating large fluctuations in pixel intensity. In contrast, at the highest speed, the values of outliers get smaller, suggesting that motion blur at higher speeds smooths out these large variations.

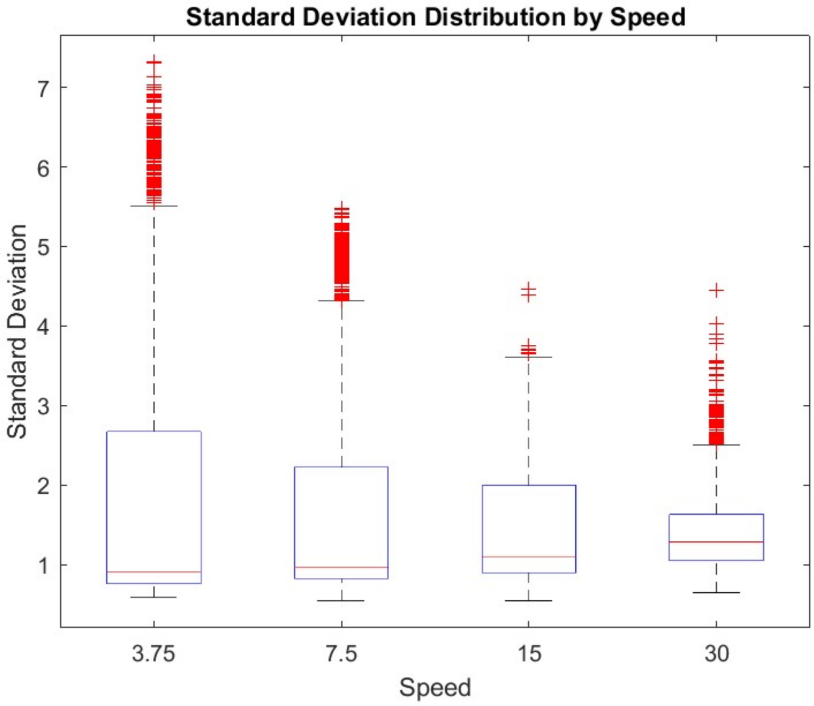


Figure 2.19: The distribution of standard deviation for different speeds.

To ensure that the observed changes in values are not random and they reflect a statistically significant relationship with speed, an ANOVA test was conducted. The results, summarized in Table 2.2, indicate significant differences in standard deviations across the various speeds (F(3, 12148) = 97.4875, p < 0.0001). This statistical finding supports the visual findings from the box plot and the descriptive statistics, and confirms that playback speed has an impact on the variability of pixel intensities.

Source	**SS**	**df**	**MS**	**F**	**Prob > F**
Groups	378.7443	3	126.2481	97.4875	2.3784e-62
Error	15732.0000	12148	1.2950	-	-
Total	16111.0000	12151	-	-	-

Table 2.2: ANOVA results for standard deviations across different playback speeds.

While the insights obtained from this test are useful, there is still a need for additional data to confirm these findings, and further investigation with new data is necessary to fully understand the effects of playback speed on pixel intensity variations.

2.6 Evaluation

To assess the improved anomaly detection method, a series of tests were performed using video recordings of open-reel tapes, each containing various types of irregularities—primarily splices, as they are the most frequent. The objective of the experiment was to evaluate the method's performance at different playback speeds by analyzing precision, which measures the system's ability to accurately identify anomalies without mistakenly labeling normal sections, and recall, which indicates how well the method detects all actual anomalies present.

The recordings were captured at four speeds: 3.75 ips, 7.5 ips, 15 ips, and 30 ips, with varying tape lengths: 10 minutes for 3.75 ips, 34 minutes for 7.5 ips, 10 minutes for 15 ips, and 9 minutes for 30 ips. Specifically, the 3.75 ips tape had 14 splices, the 7.5 ips tape included 66 splices, 4 annotations(considered shadows), 3 shadows, and 3 end-of-tape markers. The 15 ips tape featured 55 splices and 1 shadows, while the 30 ips tape contained 95 splices and 2 annotations.

Table 2.3 outlines the results for each playback speed, demonstrating the method's effectiveness across all speeds. Although it achieved perfect recall by identifying every anomaly, it did generate some false positives.

Speed (ips)	**Precision**	**Recall**
3.75	0.9333	1.0000
7.5	0.9268	1.0000
15	0.9655	1.0000
30	0.8818	1.0000

Table 2.3: Precision and recall of the new anomaly detection method at different playback speeds.

While the new method detected all irregularities at the tested speeds, there were occasional false positives. This issue was particularly noticeable at the 30 ips speed, where the motion blur caused by faster playback made it necessary to lower the detection threshold, thereby increasing the system's sensitivity to frame differences. This adjustment was made to compensate for the effects of motion blur but resulted in more false positives, especially due to lighting variations and vibrations. Additionally, some of the false positives were duplicates of the same anomaly. The design of the new method prioritizes capturing anomalies that occur in close succession, which can sometimes lead to duplicate detections. Future improvements could address this by merging consecutive anomalies into a single detection after the classification phase, reducing false positives and improving precision.

The method's focus on analyzing intensity variations between frames, combined with its use of advanced filtering techniques, has shown promising results in detecting anomalies across different speeds and video conditions. However, further refinements are planned to enhance its robustness and accuracy.

Chapter 3

Dataset Construction

Aside from detecting the existence of irregularities, Video Analyser module is also tasked for providing images for Tape Irregularity Classifier [15]. For proper functionality of the classifier, a dataset should be constructed by addressing two important changes that the program undergoes: change of video properties and recently employed frame differencing method.

Firstly, with the acquisition of new equipment by CSC, it is now feasible to record videos with better resolution and frame rate. Currently, the process of expanding the video archive with better quality videos is being carried out. In addition to the existing interlaced videos with PAL (720x576) resolution and 25 frames per second, new progressive videos are being recorded with a resolution of 1920x1080 and 50 frames per second. This change in quality affects the content of the images. Therefore it is crucial to construct the dataset accordingly.

Secondly, with the application of the aforementioned frame differencing method, it is essential to train the classifier module with the new output images. These images should adhere to the adapted method so that classifier functions properly.

3.1 Dataset Content

Following the frame differencing method, the images are categorized in four sets, each representing a different stage of irregularity detection process. The reasoning for organizing these groups stems from the fact that each stage possesses varying degrees of complexity. The objective is to identify the group that offers the most valuable and insightful representation of the irregularity. The sets are as follows:

1. Region of Interest Images: Images of ROI directly taken from the video without any processing step.
2. Difference Images: The image that represents the absolute difference between two frames at the given time.
3. Motion Images: Binarized images obtained by applying threshold on difference images.
4. Final Images: The images after the opening operation on motion images.

Each set comprises three subsets in accordance with the guidelines established by CSC. All irregularity images are categorized in these three groups: splices, shadows and brands.

An important point to note is that every image in one of these four sets corresponds to its counterpart in another set, as they are products of the same process. Therefore four groups have the same number of images representing the same irregularities. The reasoning is to attain more reliable and consistent empirical results.

Similar to the irregularity detection process, the dataset is constructed by using videos with speed options 3.75 ips, 7.5 ips, 15 ips and 30 ips. Inclusion of 30 ips is crucial because higher tape speed introduces new challenges for irregularity classification due to the increase of motion blur. In the future iterations, the dataset should be expanded by obtaining images with 1.875 ips and 0.9375 ips that exists in ARP standard [15].

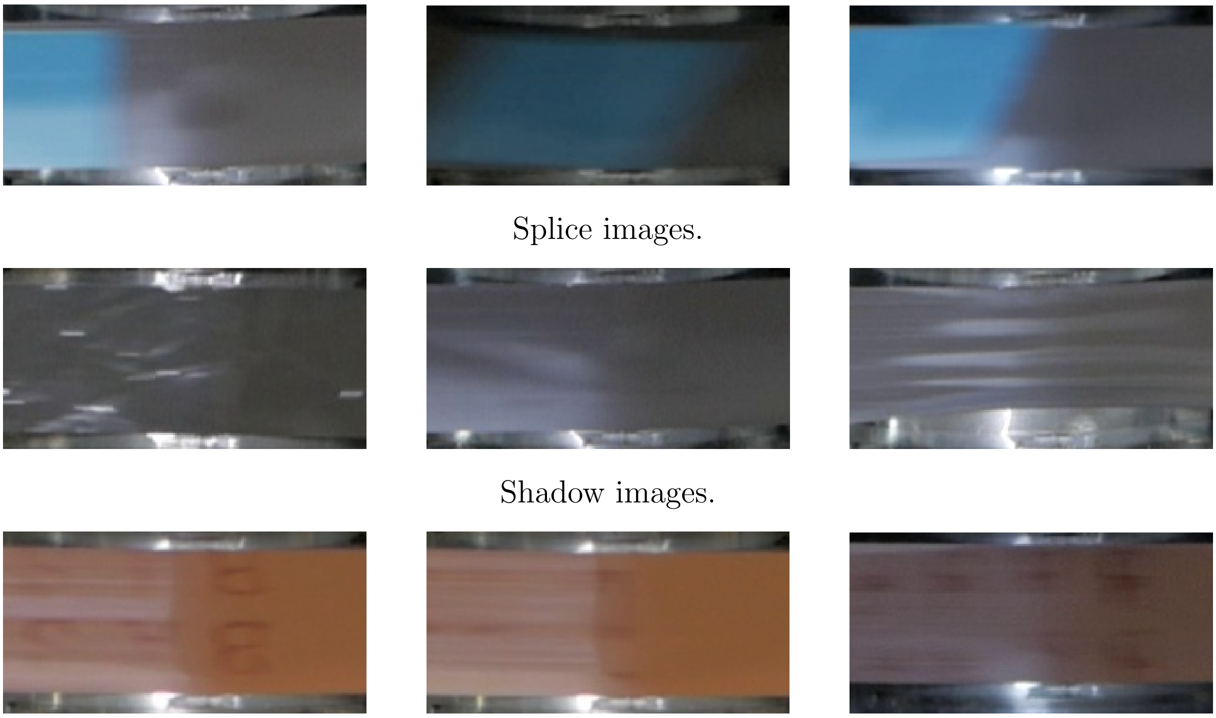
3.1.1 Region of Interest Images

First set of images consists of direct images taken from the tape area of the video. These images represent the second image of the consecutive pairs of frames from the detection process.

In Fig 3.1, the some typical images of this category are shown.

3.1.2 Difference Images

These images represent the difference of two consecutive frames at the time the irregularity was detected. These images are obtained by finding the absolute difference between pixel intensities of two consecutive images as discussed earlier in Section 2.2.2. Since the pixel intensities of these difference images are relatively low, they predominantly appear black to the human eye. However these images may carry meaningful components discernible to machine vision.



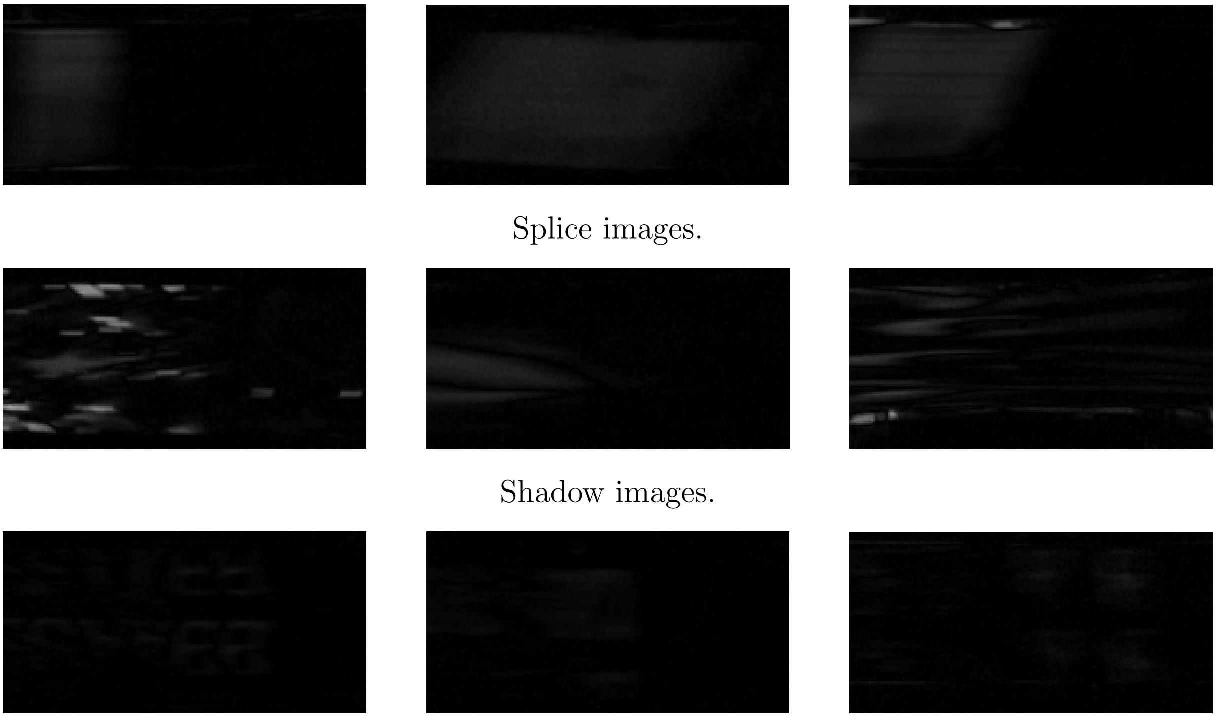
Splice images.

Shadow images.

Brand images.

Figure 3.1: ROI images with various speeds.

In Fig 3.2, some images of this category are shown. These images are obtained by finding the absolute difference for the irregularities in Fig 3.1. It can be seen that brand images appear mostly black, attributed to their typically low intensities.



Splice images.

Shadow images.

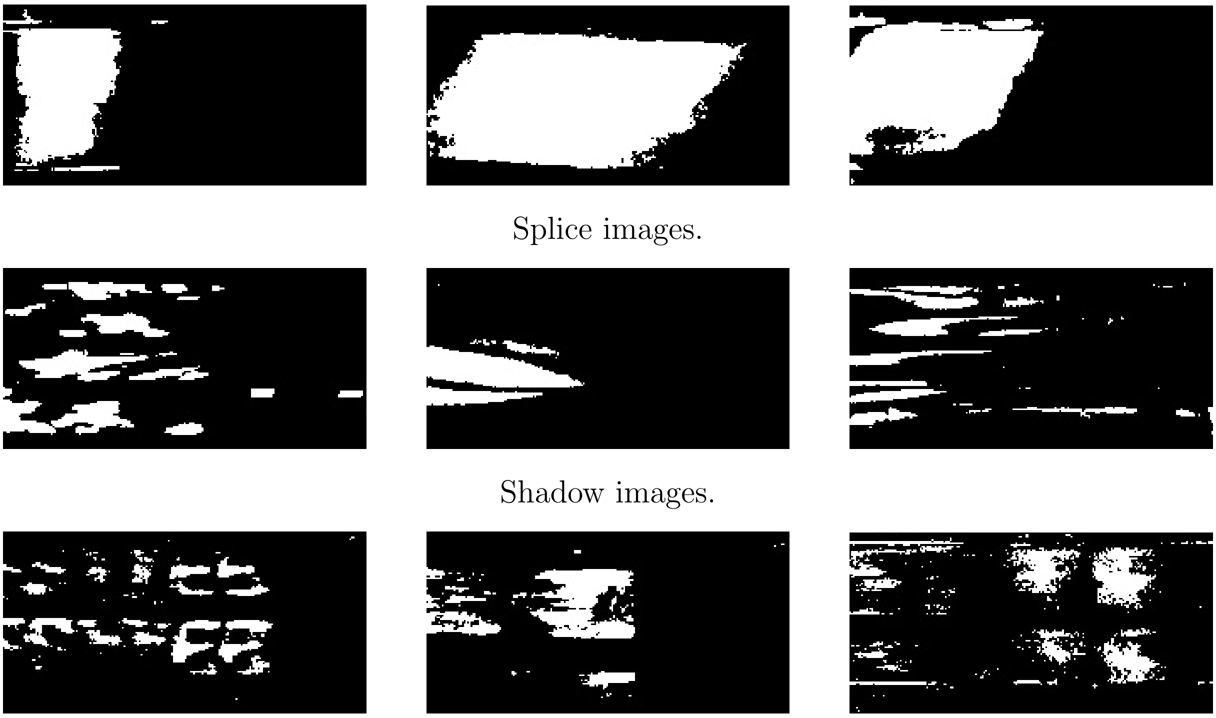
Brand images.

Figure 3.2: Difference images with various speeds.

3.1.3 Motion Images

Motion images are the images obtained by binarizing a difference image based on the obtained threshold value. The thresholding procedure is employed as discussed earlier in Section 2.2.5.

In Fig 3.3, some motion images are shown. These images are the binarized versions of the the images in Fig 3.2 with optimal threshold values.



Splice images.

Shadow images.

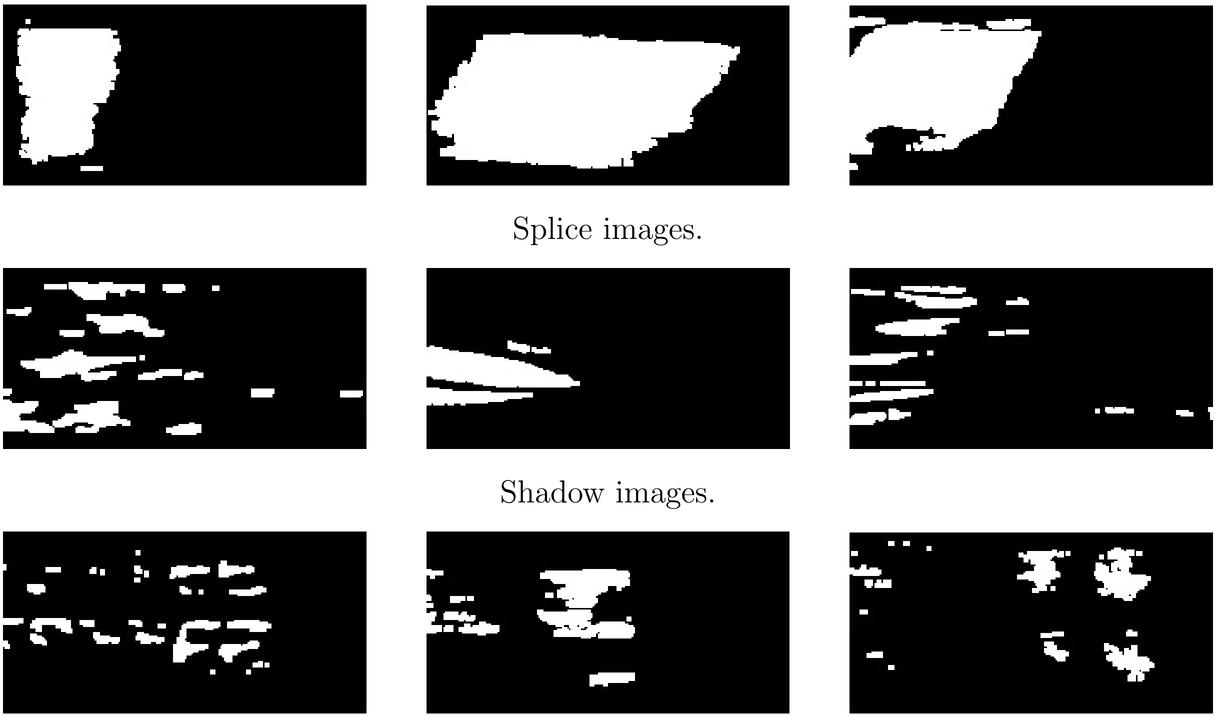
Brand images.

Figure 3.3: Binarized motion images with various speeds.

3.1.4 Final Images

Final images are the simplest representations of irregularities and are the images obtained by employing opening operation on motion images. Opening operation is implemented as mentioned earlier in Section 2.2.6. The reasoning behind creating a set of these images is that, by destroying the undesirable motion artifacts on motion images, it may be possible to get a simpler and more accurate representation of the irregularity that is interpretable by machine.

In Fig 3.4, some final images are shown. These images are obtained by applying opening operation on image in Fig 3.3.



Splice images.

Shadow images.

Brand images.

Figure 3.4: Opened final images with various speeds.

3.2 Handling Old-Video Dataset

Since the video archive of CSC predates the development of MPAI project, the archive features a considerable amount of interlaced videos [19]. Due to the availability of equipment at the time, interlacing was deemed necessary. Since recording the entire archive with better equipment is not feasible, it is crucial to repurpose these old videos and address the unique challenges they introduce.

3.2.1 Interlacing

Interlacing is a technique used in video transmission and display where a single frame is split into two separate fields: an odd field containing all the odd-numbered lines and an even field containing all the even-numbered lines. These fields are transmitted or displayed alternately, with the odd field displayed first followed by the even field. Although this method was primarily utilized with cathode ray tube (CRT) displays and is associated with legacy systems, it can still be employed due to the bandwidth and storage advantages it offers [26]. The interlacing can be expressed as follows:

$$\text{Interlaced Frame}[i,j] = \begin{cases} \text{Field 1}[i,j] & \text{if } i \text{ is even} \\ \text{Field 2}[i,j] & \text{if } i \text{ is odd} \end{cases}$$

In interlaced video frames, the lines of motion may appear misaligned due to the temporal offset between odd and even fields. This occurs because each field captures the scene at a slightly different point in time, resulting in a discontinuity in the motion between the two fields.

In Fig 3.5, the misaligned lines are visible in tape area.

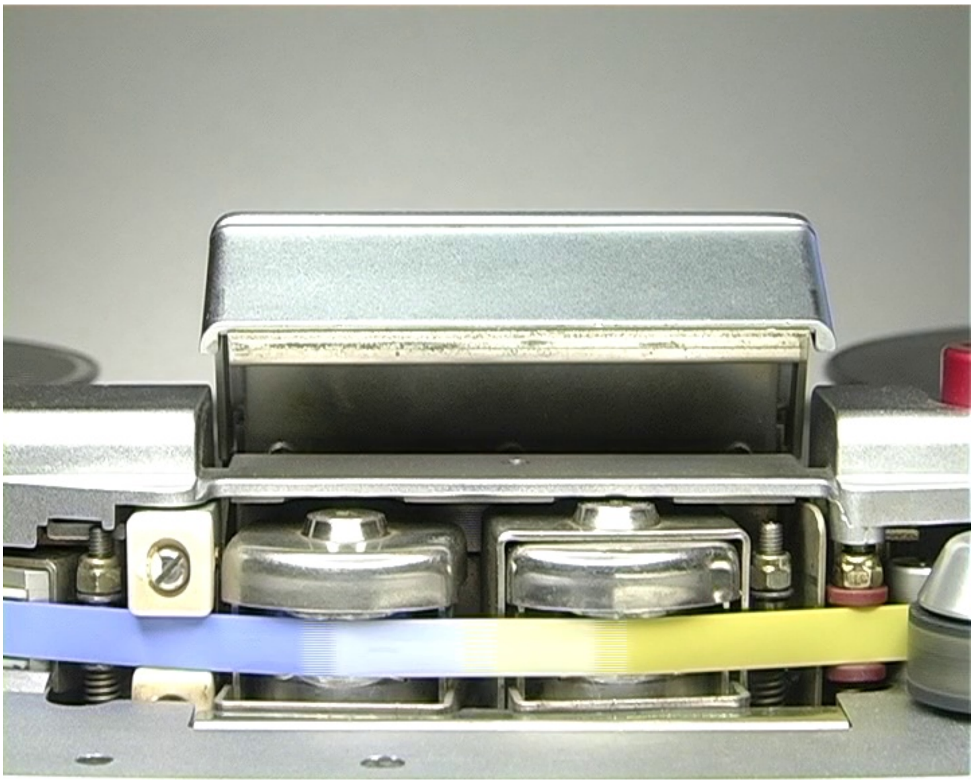
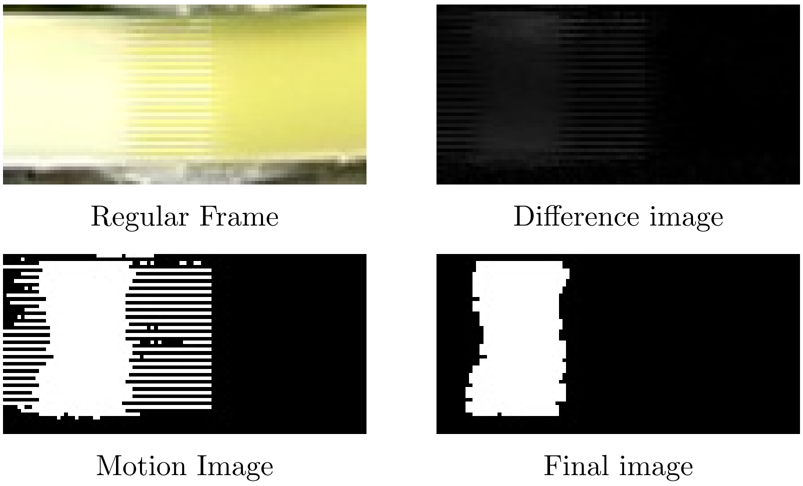


Figure 3.5: An interlaced image.

These misalignment lines pose a unique challenge as the output images of frame differencing method fail to preserve the integrity of the irregularity. For the classifier to function optimally, it is important to mitigate the effects of interlacing by employing an effective deinterlacing method.

In Fig 3.6, the effects of interlacing can be seen. It is the same moments in the video as 3.5.



Regular Frame

Difference image

Motion Image

Final image

Figure 3.6: Frame differencing process applied on an interlaced video

3.2.2 Interpolation-Based Deinterlacing

Deinterlacing is a process used to convert interlaced video, which consists of alternating fields of odd and even lines, into progressive video. Some common deinterlacing methods include bob, weave, and interpolation-based techniques [27].

The interpolation-based deinterlacing method aims to generate new pixel values for the missing lines in each field using interpolation techniques. One common interpolation method is linear interpolation, which calculates the intermediate pixel values based on the known pixel values in adjacent lines [28].

Let P_1 denote the pixel value at the left or lower known position, x_1, in the image. Let P_2 represent the pixel value at the right or upper known position, x_2, in the image. The coordinates x_1 and x_2 correspond to the positions of the known pixel values P_1 and P_2. Consider the coordinate x at which the interpolation of the pixel value is desired within the image. Based on the known pixel values P_1 and P_2 and their respective positions x_1 and x_2, the interpolated pixel value $\bar{P}(x)$ can be estimated at coordinate x. The linear interpolation formula is given by:

$$\bar{P}(x) = P_1 + (x - x_1) \cdot \frac{P_2 - P_1}{x_2 - x_1}$$

This formula calculates the value at point x by linearly interpolating between the known pixel values P_1 and P_2. The result is a smooth transition between the two known values.

The interpolation process involves the following steps:

1. For each pair of adjacent odd rows, calculate the average of corresponding pixel

values to generate new pixel values for the missing even row.

2. Repeat this process for all pairs of adjacent odd rows in the image.

3. Fill in the missing lines between the odd rows using the interpolated pixel values, effectively increasing the vertical resolution of the deinterlaced frame and producing smoother transitions between rows.

In Fig 3.7, the generated frame after employing deinterlacing on Fig 3.5 is shown. The misalignment lines are largely diminished.

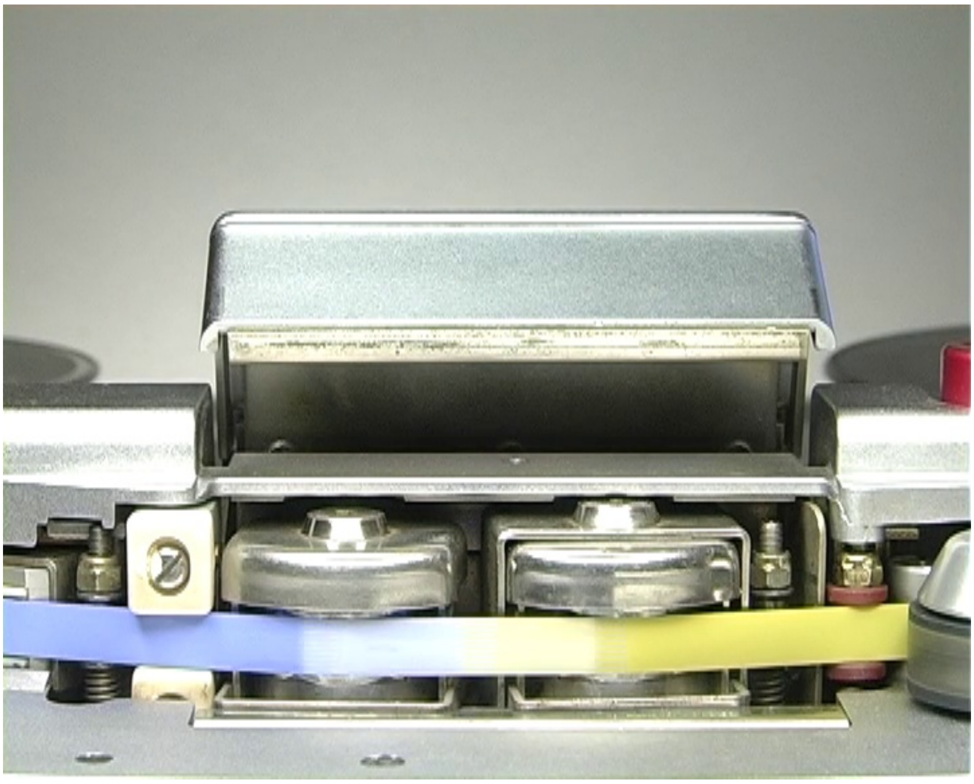
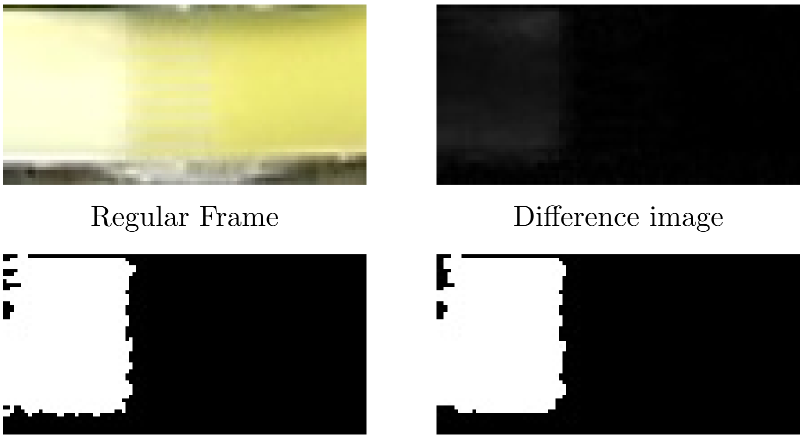


Figure 3.7: Deinterlaced image of Fig 3.5.

In Fig 3.8, the resulting images of frame differencing are shown after deinterlacing. It can be observed that the integrity of the irregularity is more effectively preserved compared to 3.6.

The interpolation-based deinterlacing method is implemented for its simplicity and effectiveness. Unlike bob, which simply duplicates fields, and weave, which merges fields without interpolation, interpolation-based deinterlacing preserves the integrity of the original image by accurately interpolating missing lines [27].

While linear interpolation serves as an effective deinterlacing method, it is important to note that there exist more advanced techniques beyond interpolation, which can be implemented in future iterations if deemed necessary.



Regular Frame

Difference image

Motion Image

Final image

Figure 3.8: Frame differencing process applied on an interlaced video

Chapter 4

Irregularity Classification

The Tape Irregularity Classifier is responsible for categorizing irregularities identified by the Audio Analyzer and Video Analyzer using a machine learning model. For visual data classification, constructed datasets were trained with a machine learning module, and classification was performed for each dataset. Testing with various models is currently ongoing. In this thesis, ResNet-50, the model demonstrating the highest success rate, is discussed.

4.1 ResNet

ResNet (Residual Network) is a type of deep learning architecture designed to improve the training of very deep networks by addressing the issues of vanishing and exploding gradients that often occur in traditional deep neural networks [29].

Gradient descent is an optimization algorithm used for minimizing the error of a model by iteratively adjusting the parameters of the model. This method calculates the gradient (i.e., the direction and rate of fastest increase) of the loss function concerning the parameters and updates the parameters in the opposite direction of the gradient to reduce the loss [30].

Gradient descent is a widely used optimization method but it introduces some challenges in neural networks with high number of layers. One such challenge is vanishing gradients. As gradients are backpropagated through the network, they can diminish exponentially. This may make it difficult to effectively update the earlier layers. The other challenge is exploding gradients. In this case, the opposite happen can happen and gradients can grow exponentially, causing updates to be unstable. This leads to poor convergence [31].

ResNet introduces a residual learning framework which allows the network to learn residual functions with reference to the layer inputs, rather than trying to learn unreferenced functions. This approach fundamentally changes how layers in a deep network interact. Instead of each layer learning a completely new representation, it learns modifications to the input layer, formulated as:

The residual function:

$$F(x) = H(x) - x$$

where $H(x)$ is the desired mapping and x is the input. This leads to the formulation:

$$H(x) = F(x) + x$$

This concept allows deeper networks to be trained effectively because it mitigates the problem of vanishing gradients by providing an easier optimization path through identity mappings [29].

The unique quality of ResNet is the use of identity shortcut connections that bypass one or more layers. These shortcuts are essentially identity mappings added directly to the output of the convolutional layers. This technique does not introduce additional parameters or computational complexity, making it highly efficient. The identity shortcuts enable the gradient to go through the network without diminishing. This preserves the learning capability of the model even at depths. The innovation of Residual Network allows the construction of networks with hundreds of layers. ResNet architectures with 50, 101, or even 152 layers have been shown to perform well [29].

The rationale of choosing ResNet is its flexible design and its capability of dealing with a wide range of image types. Its versatile and robust architecture makes it suitable for both natural images with complex pixel intensities such as ROI images and simple binary images such as motion images.

For this research, ResNet-50 has been used, yielding the best results so far. However, tests with other models, such as YOLO [32], a real-time object detection system predicting bounding boxes and class probabilities in one evaluation, and VGG [33], a deep convolutional neural network using small filters to improve image recognition, are currently being conducted.

4.2 Expanding the Dataset

Despite the extensive number of magnetic tapes under the disposal of CSC, the presence and distribution of irregularities are unorganized. Therefore the dataset, as it is, is not comprehensive of the real world scenarios and this causes the data to be incomplete.

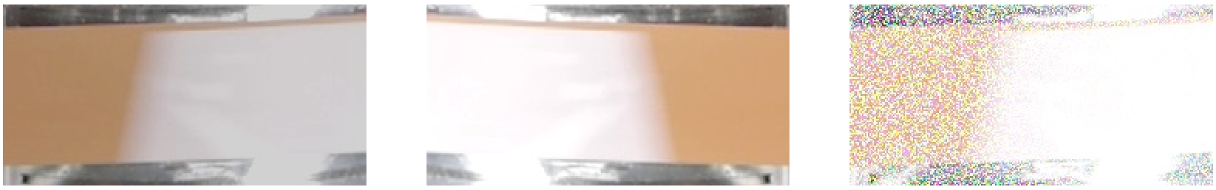
To address this challenge, data augmentation is applied to expand the dataset. Data augmentation involves increasing the number of images in the dataset by creating modified copies through various transformation techniques [30]. The purpose of this operation is to prevent overfitting and to enhance the model's ability to generalize classification. Simple augmentation techniques include geometric transformations such as horizontal and vertical flipping, scaling, and rotating, as well as color space adjustments like modifying brightness, contrast, and saturation. It should be noted that data augmentation techniques extend beyond basic operations to include more complex methods. These methods, including random alterations of color values [34], complex geometric distortions of input images [35], and elastic distortions [36], are widely used and have demonstrated effective results.

Due to the diverse characteristics of irregularities within the "shadow" class and the variable tape speed options, data augmentation techniques were restricted to a select few. Selected augmentation techniques were implemented using the imgaug library [37], incorporating the following methods:

- Flipping: Images are horizontally or vertically flipped to increase variation in the dataset, helping the model generalize better to unseen data.
- Gaussian Noise: Random noise with a Gaussian distribution, where the scale parameter ranges from 0 to 0.05 times 255, is added to images to enhance robustness of the model by simulating real-world variability in image data.
- Brightness Change: The brightness of images is randomly adjusted within a range of plus or minus 20%. which diversifying the dataset and improving the model's ability to handle varying levels of lighting.

In Fig 4.1, some sample images created by using aforementioned operations on a splice image are shown.

Initially, scaling and rotating were considered for inclusion but were omitted due to their divergence from real-world scenarios, due to the strictly defined borders of the region



(a) 82% Brightness (b) Horizontal Flip (c) Gaussian Noise

Figure 4.1: Sample images of a splice produced by augmentation.

of interest. Distortion methods were also excluded because the irregularities within the ’shadow’ class lack a defined shape.

4.3 Classification Results

After the expansion, these four datasets are trained using ResNet-50 architecture for classification. The following sections summarize the classification results for each dataset, highlighting the effectiveness and limitations of each approach. ”Since the focus of this research is on comparing the results, the resulting values are rounded to two decimal places for simplicity.

4.3.1 Results for ROI Images

The results for ROI images show exceptional performance across all classes. For the ”brand” class, the precision is 0.99, recall is 1.00, and F1-score is 1.00, indicating near-perfect classification with almost no false positives or false negatives. Similarly, the ”shadow” and ”splice” classes also exhibit high metrics with precision and recall both near 0.98, resulting in an F1-score of 0.98.

The high precision and recall across all classes show the model’s capability to correctly identify and classify the ROI images. The consistency in high F1-scores is an indicator of the robustness of the model when dealing with natural images of regions of interest.

In Fig 4.2 and Table 4.1, the results for ROI images and confusion matrix are shown.

	precision	recall	f1-score
brand	0.99	1.00	1.00
shadow	0.98	0.98	0.98
splice	0.98	0.97	0.98

Table 4.1: Results for the ROI images

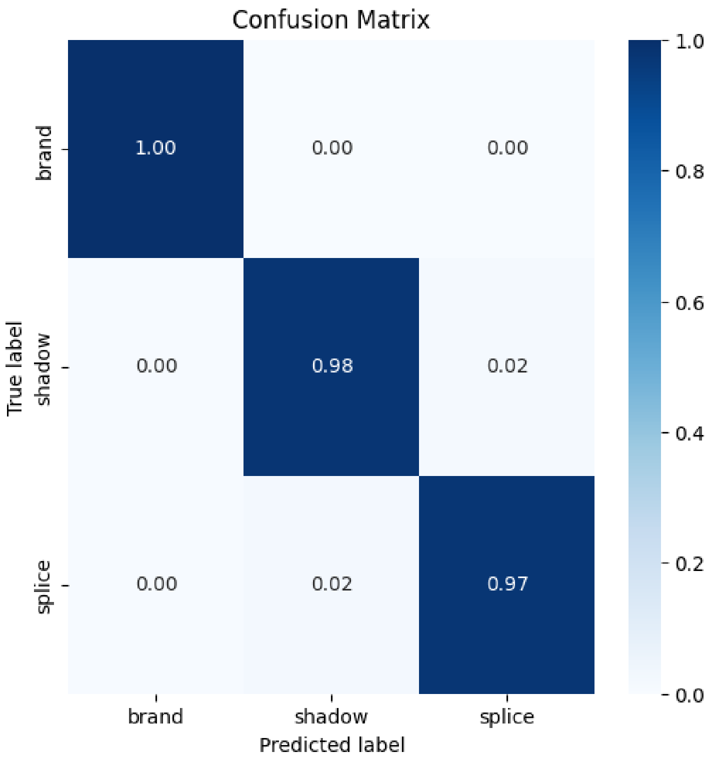


Figure 4.2: Normalized confusion matrix of natural ROI images.

4.3.2 Results for Difference Images

The results for difference images, which involve comparing frame differences to identify changes, show a slight drop in performance compared to ROI images. The ”brand” class has a precision of 0.94, recall of 0.89, and an F1-score of 0.92. These scores indicate some challenges in accurately detecting this class. This may be caused by the increased variability or noise in the difference images. The ”shadow” class fares better with a precision of 0.91 and recall of 0.95, resulting in an F1-score of 0.93. The ”splice” class maintains a balanced performance with precision and recall both around 0.91.

Despite the drop in performance, the results still indicate a reasonably good classification capability. The lower recall for the ”brand” class suggests that the model misses some instances of this class in the difference images. However, the overall F1-scores being above 0.90 for all classes show that the model can effectively classify difference images.

In Fig 4.3 and Table 4.2, the results for difference images and confusion matrix are shown.

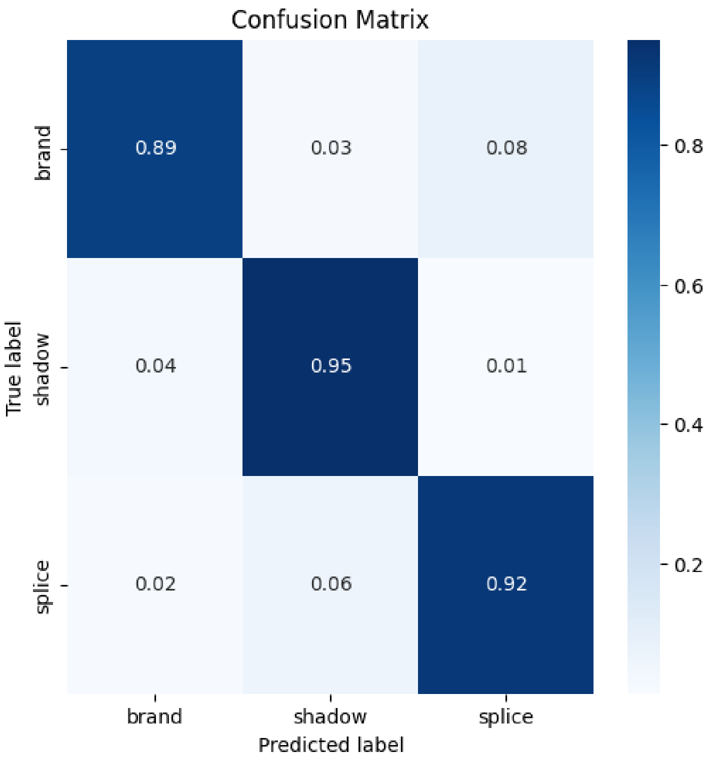


Figure 4.3: Normalized confusion matrix of difference images.

	precision	recall	f1-score
brand	0.94	0.89	0.92
shadow	0.91	0.95	0.93
splice	0.91	0.92	0.91

Table 4.2: Results for the motion images

4.3.3 Results for Motion Images

The results for motion images, which are the binarized representations of difference images, show an improvement over difference images. The ”brand” class achieves a precision of 0.98, recall of 0.92, and an F1-score of 0.96, indicating better detection accuracy and fewer false positives compared to the previous case. The ”shadow” class shows good performance with a precision of 0.94 and recall of 0.97, leading to an F1-score of 0.95. The ”splice” class exhibits a precision of 0.96, recall of 0.97, and an F1-score of 0.96, reflecting accurate and consistent classification.

The higher precision and recall values for motion images compared to difference images suggest that the binary nature of the images simplifies the classification task. This improvement indicates that motion images, despite being simpler, provide more clear and more easily distinguishable features. Therefore, it can be said that the preprocessing steps that involve thresholding yields good results.

In Fig 4.4 and Table 4.3, the results for motion images and confusion matrix are shown.

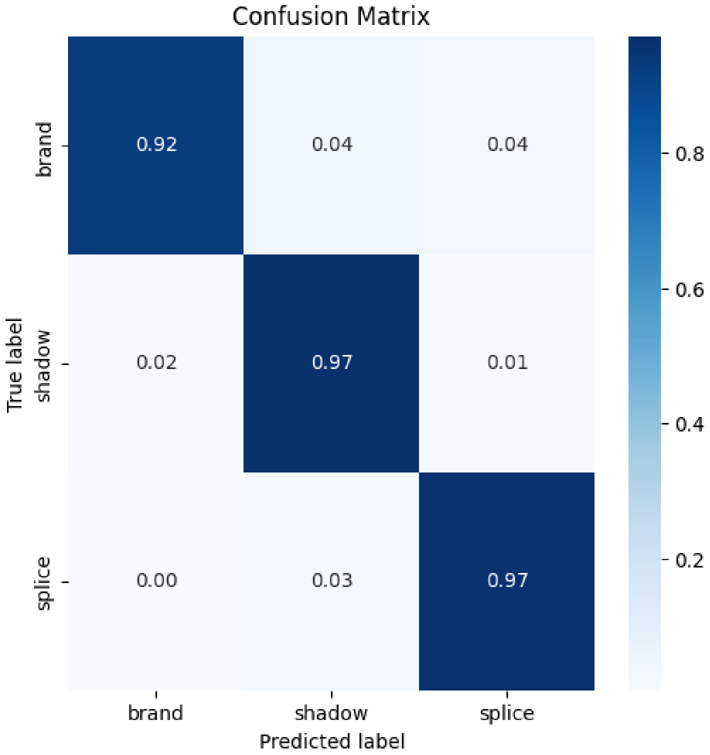


Figure 4.4: Normalized confusion matrix of motion images.

	precision	recall	f1-score
brand	0.98	0.92	0.96
shadow	0.94	0.97	0.95
splice	0.96	0.97	0.96

Table 4.3: Results for the motion images

4.3.4 Results for Final Images

The results for the final images, which involve further processing such as opening operations, show a mixed performance. The ”brand” class has a precision of 0.96 but a lower recall of 0.85, resulting in an F1-score of 0.90. This indicates that the model is good at identifying true positives but it misses a significant number of actual instances. The ”shadow” and ”splice” classes perform better, with both achieving an F1-score of 0.94, reflecting high precision and recall.

The results for final images display the trade-offs involved in preprocessing steps. The lower recall for the ”brand” class suggests that some true instances are being filtered out and not detected due to the opening operation. Therefore it is observed that opening operation can remove not only small noise, but also potentially important features. Nevertheless, the overall performance remains relatively good. This indicates that the final preprocessing step helps in refining the classification but introduces some challenges that need to be addressed, especially for the ”brand” class.

In Fig 4.5 and Table 4.4, the results for final images and confusion matrix are shown.

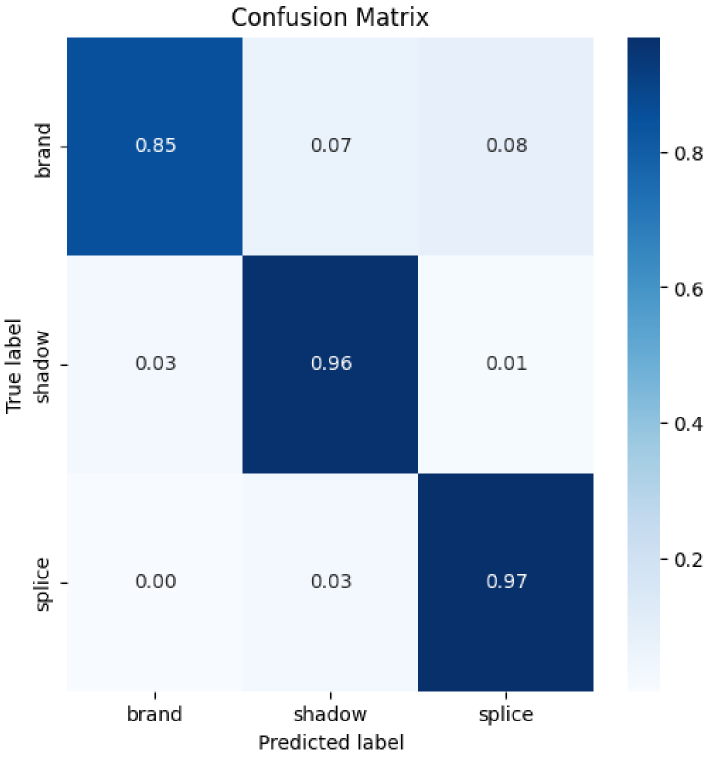


Figure 4.5: Normalized confusion matrix of final images.

	precision	recall	f1-score
brand	0.96	0.85	0.90
shadow	0.91	0.96	0.94
splice	0.91	0.97	0.94

Table 4.4: Results for the final images

4.3.5 Discussion

The classification model showed different levels of accuracy and reliability with various image processing techniques. Natural images focusing on ROI performed the best. They had nearly perfect precision, recall, and F1-scores for all categories, showing strong detection and classification abilities. However, Difference images did not perform as well, especially for the ”brand” category. This suggests challenges in handling variability or noise from frame differencing. Motion images, which included simplified binary representations, performed better than Difference images. They showed high precision and recall, benefiting from their straightforward features. On the other hand, Final images processed with opening operations had strong but somewhat inconsistent results. There was a noticeable drop in recall for the "brand" category. This indicates opening operation removes the distinguishable features of images in "brand" class. Overall, these results highlight the model's ability to handle different image types.

Chapter 5

Conclusion

The promising classification results obtained from motion images demonstrate the effectiveness of the frame differencing methodology employed in this study. However, the lower performance observed with final images suggests that the inclusion of the images produced by opening operation in the classification phase may not be beneficial and it should be reconsidered.

So far, natural ROI images yield the highest classification accuracy. This suggests that this method may be the most suitable approach for classification. However, the research data is limited because irregularities occur with varying frequencies on different tapes. Therefore, further experimentation with more extensive datasets is recommended to ensure reliability and applicability in various situations.

If ROI images consistently perform better than motion images, future research may explore relegating frame differencing only to detection tasks rather than integrating it into the classification process.

Even if frame differencing is removed from classification, the method remains useful for visualizing irregularities and providing insights that can guide how data is gathered in the future. It serves as a complementary tool for human inspection and analysis.

5.1 Future Work

5.1.1 Inclusion of Remaining Speed Options

In this thesis, the focus was on four speed options: 3.75 ips, 7.5 ips, 15 ips, and 30 ips. However, it is noted that there are other speed options such as 0.9375 ips and 1.875 ips [15]. Future research should extend the methodology to include observation and analysis of tapes with these speeds. Adopting appropriate methodologies for these speeds will enhance the comprehensiveness of the research.

5.1.2 Expanding the Classification

Increasing the Number of Classes

In this thesis, the classification was done using three predefined classes, consistent with previous research standards. However, using unsupervised learning methods like clustering could help discover new classes based on visual similarities among images. This approach could improve classification and offer a better representation of the data.

False Detection

The detection system is susceptible to external factors. Rapid changes in light or vibrations can trigger it. While static images may not reveal anomalies, processed images often show easily recognizable shapes. However, the dataset for these unexpected situations was limited. In the future, experiments could be conducted in controlled environments, integrating these images into the classification process.

5.1.3 Concluding the Future Work

To conclude, by addressing additional speed options and expanding classification methods, future research aims to strengthen the reliability and applicability of the framework of automated irregularity detection for audio preservation. These advancements seek to refine classification accuracy and to advance the management of irregularities in magnetic audio tape preservation

Bibliography

[1] Claudia Fantozzi, Federica Bressan, Niccolò Pretto, et al. Tape music archives: from preservation to access. *Int J Digit Libr*, 18(3):233–249, 2017.

[2] Niccolò Pretto, Carlo Fantozzi, Edoardo Micheloni, Valentina Burini, and Sergio Canazza. Computing methodologies supporting the preservation of electroacoustic music from analog magnetic tape. *Computer Music Journal*, 42(4):pp. 59–74, 2018.

[3] Sergio Canazza, Giovanni De Poli, and Alberto Vidolin. Gesture, music and computer: the Centro di Sonologia Computazionale at Padova University, a 50-year history. *Sensors*, 22(9), 2022.

[4] Marina Bosi, Sergio Canazza, Alessandro Russo, Niccolo Pretto, and Leonardo Chiariglione. An MPAI/IEEE international standard for audio: Overview of CAE audio recording preservation (ARP) technology. In *Audio Engineering Society Conference: 2023 AES International Conference on Audio Archiving, Preservation & Restoration*, Jun 2023.

[5] Federica Bressan and Sergio Canazza. A systemic approach to the preservation of audio documents: Methodology and software tools. *Journal of Electrical and Computer Engineering*, 2013:1–16, 2013.

[6] International Federation of Library Associations and Institutions (IFLA). *Guidelines for Audiovisual and Multimedia Materials in Libraries and Other Institutions.* International Federation of Library Associations and Institutions, 2004.

[7] I. T. Committee, International Association of Sound and Audiovisual Archives. *IASA-TC 03 The Safeguarding of the Audiovisual Heritage: Ethics, Principles and Preservation Strategy.* International Association of Sound and Audiovisual Archives, 4th edition, 2017.

[8] Kevin Bradley. *IASA TC-04 Guidelines in the Production and Preservation of Digital Audio Objects*. International Association of Sound and Audiovisual Archives, 2nd edition, 2009.

[9] Sergio Canazza Targon and Giovanni De Poli. Four decades of music research, creation, and education at padua's centro di sonologia computazionale. *Computer Music Journal*, 43(4):58–80, 2020. Open Access.

[10] MPAI. About mpai. https://mpai.community/about/. Accessed: 22-05-2024.

[11] Marina Bosi, Niccolò Pretto, Marco Guarise, and Sergio Canazza. Sound and music computing using ai: designing a standard. In *Proceedings of the 18th Sound and Music Computing Conference, SMC 2021*, pages 215–218, Virtual Conference, 2021. Sound and Music Computing Network. International conference.

[12] IEEE Standards Association. IEEE Standard Adoption of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Technical Specification Artificial Intelligence Framework (AIF) 1.1, April 2023.

[13] Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI). Technical Specification: AI Framework (MPAI-AIF) Version 2. Technical report, MPAI, 2023.

[14] IEEE. IEEE standard adoption of moving picture, audio and data coding by artificial intelligence (mpai) technical specification context-based audio enhanced (cae) version 1.4, 2023.

[15] Context-based Audio Enhancement MPAI-CAE. Technical report, MPAI Community, 2023. Version 2.1.

[16] Sergio Canazza, Emery Schubert, Anna Chmiel, Nicola Pretto, and Antonio Rodà. The magnetic urtext: Restoration as music interpretation. *Frontiers in Psychology*, 13:844009, 2022.

[17] Nicola Orio, Lauro Snidaro, Sergio Canazza, and Gian Luca Foresti. Methodologies and tools for audio digital archives. *International Journal on Digital Libraries*, 10(4):201–220, 2009.

[18] Mary Miliano, editor. *The IASA Cataloguing Rules*. International Association of Sound and Audiovisual Archives, London, 1999.

[19] Alessandro Russo, Matteo Spanio, and Sergio Canazza. Enhancing preservation and restoration of open reel audio tapes through computer vision. pages 297–308, 2024.

[20] A.J. Lipton, H. Fujiyoshi, and R.S. Patil. Moving target classification and tracking from real-time video. pages 8–14, 1998.

[21] Dana H Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981.

[22] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[23] OpenCV Development Team. *OpenCV 2 Documentation*. OpenCV, 2011.

[24] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[25] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2nd edition, 2022.

[26] A. Murat Tekalp. *Digital Video Processing*. Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 1995.

[27] G. De Haan and E.B. Bellers. Deinterlacing-an overview. *Proceedings of the IEEE*, 86(9):1839–1857, 1998.

[28] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2008.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[30] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[31] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[35] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[36] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Best practices for convolutional neural networks applied to visual document analysis. *International Conference on Document Analysis and Recognition*, pages 958–962, 2003.

[37] Alexander Buslaev, Alex Parinov, Vladimir Iglovikov, and Eugeny Khvedchenya. imgaug documentation. https://imgaug.readthedocs.io/, Accessed: 13-06-2024.