TU WIEN Informatics

Shot Boundary Detection: Eine grundlegende Basis für die automatische Videoanalyse

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science

im Rahmen des Studiums

Logic and Computation

eingereicht von

Deana Zafirova, BSc
Matrikelnummer 01227664

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Privatdoz. Dipl.-Ing. Dr.techn. Martin Kampel
Mitwirkung: Projektass. MSc Daniel Helm

Wien, 9. Oktober 2020

Deana Zafirova	Martin Kampel

TU WIEN Informatics

Shot Boundary Detection: A Fundamental Base for Automatic Video Analysis

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Logic and Computation

by

Deana Zafirova, BSc
Registration Number 01227664

to the Faculty of Informatics

at the TU Wien

Advisor: Privatdoz. Dipl.-Ing. Dr.techn. Martin Kampel
Assistance: Projektass. MSc Daniel Helm

Vienna, 9th October, 2020

Deana Zafirova	Martin Kampel

Erklärung zur Verfassung der Arbeit

Deana Zafirova, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 9. Oktober 2020

Deana Zafirova

Danksagung

Ich möchte mich ganz herzlich bei meinem Betreuer Prof. Kampel für seine Betreuung und Anleitung bedanken. Außerdem möchte ich mich bei Projektass. Helm für die Beratung und das geteilte Wissen während des Prozesses der Arbeit bedanken. Abschließend möchte ich meiner Familie und Stefan für ihre uneingeschränkte Geduld, Unterstützung und Motivation danken.

Acknowledgements

I would like to express my gratitude and especially thank my supervisor Prof. Kampel for his mentorship and guidance. Further, I would like to thank Projektass. Helm for the advisory and shared knowledge during the process of the thesis. Finally, I wish to thank my family and Stefan for their unlimited patience, support and motivation.

Kurzfassung

Die Shot Boundary Detection (SBD) ist ein grundlegender Vorverarbeitungsschritt bei der automatisierten inhaltsbasierten Videoindizierung und -abfrage (CBVIR). Verwandte Forschungsarbeiten konzentrieren sich auf die Erkennung von abrupten Übergängen (AT) und allmählichen Übergängen (GT) in Videodatensätzen vielseitiger Domänen und Videogattungen wie Filmen, Dokumentationen und Nachrichtenclips. Allerdings widmet sich nur eine Minderheit der SBD-Forschungsstudien dem historischen Filmmaterial. Das Hauptziel dieser Masterarbeit ist es, das Problem der SBD in historischen Filmen anzusprechen. Daher wird ein neuartiges SBD-Framework vorgeschlagen, das auf einem tiefen neuronalen Netzwerk namens ResidualATNet basiert. Die grundlegende Architektur von ResidualATNet basiert einem siamesischen Netzwerk und nutzt die Cosinus-Ähnlichkeit. Das Framework umfasst einen AT-Detektor und einen GT-Detektor, die separat auf Übergänge zum Auflösen und Ausblenden (FOI) / Wischen abzielen. Zusätzlich wird ein selbst erstellter historischer Datensatz für das Training von ResidualATNet erstellt. Experimente, die die Auswirkungen der Trainingsdaten, Merkmalsextraktionsstrategien und CNN-Architektureigenschaften untersuchen, werden durchgeführt, um die SBD-Leistung auf historischen Daten zu verbessern. Die Auswertung wird an zwei historischen Datensätzen durchgeführt, die als EFilms und IMC bezeichnet werden und 66 bzw. 78 Filme enthalten. Das Framework erreicht einen F1-Score von 85% in den EFilms und einen F1-Score von 91% im IMC-Datensatz. Experimente mit den öffentlich verfügbaren Datensätzen RAI, ClipShots und BBC Planet Earth bestätigen, dass das Framework eine wettbewerbsfähige SBD-Leistung für zeitgenössisches Filmmaterial liefert. Mit einem F1-Score von 96% im RAI-Datensatz und einem F1-Score von 90% im BBC Planet Earth-Datensatz zeigt das Framework hervorragende Erkennungsfähigkeiten, die nicht auf historisches Filmmaterial beschränkt sind. Letztendlich tragen diese These und ihre Ergebnisse wesentlich zum Ziel bei, eine intelligente CBVIR-Anwendung für historische Filme zu entwickeln.

Abstract

Shot Boundary Detection (SBD) represents a fundamental preprocessing step in automated content-based video indexing and retrieval (CBVIR). Related research focuses on detecting Abrupt Transitions (AT) and Gradual Transitions (GT) in video datasets of versatile domains and video genres such as movies, documentaries and news clips. However, only a minority of the SBD research studies dedicate their attention to historical footage. The main aim of this master thesis is to address the problem of SBD in historical films. Therefore, a novel SBD framework based on a deep neural network called ResidualATNet is proposed. The basic architecture of ResidualATNet resembles a Siamese network and utilizes Cosine similarity. The framework includes an AT detector and a GT detector which targets Dissolve and Fade Out-In (FOI) / Wipe transitions separately. Additionally, a self-designed historical dataset is created for the training of ResidualATNet. Experiments which examine the effects of the training data, feature extraction strategies and CNN architectural properties are carried out to improve the SBD performance on historical data. The evaluation is performed on two historical datasets called EFilms and IMC which contain 66 and 78 films respectively. The framework achieves an F1-score of 85% on the EFilms and an F1-score of 91% on the IMC dataset. Experiments on the publicly available datasets RAI, ClipShots and BBC Planet Earth confirm that the framework produces a competitive SBD performance on contemporary film material. With an F1-score of 96% on the RAI dataset and an F1-score of 90% on the BBC Planet Earth dataset, the framework shows outstanding detection abilities which are not limited to historical film material. Ultimately, this thesis and its results significantly contribute to the goal of developing a smart CBVIR application for historical films.

Contents

CHAPTER 1

Introduction

1.1 Motivation

Videos capture audiovisual information and as such play a crucial role in our culture [SZMB11]. Historical films provide priceless insight into the past and a chance to look at our history. This unique way of documentation through sound recordings and moving images makes audiovisual heritage invaluable and vitally important. Nowadays, museums and specialised archives such as the National Archives and Records Administration (NARA) [1] and the Austrian Film Museum [2] are responsible for the storage and preservation of a large number of historical films. Furthermore, methods for automated content-based video analysis, indexing and retrieval play a crucial part in the conservation, access and search of historical film material [ZMZB11]

Content-based video indexing and retrieval (CBVIR) tools aim to identify meaningful composition structures for extraction and representation of the content of different video sources [ARS+18]. CBVIR tools are responsible for automated parsing of videos and facilitate easy accessibility, fast search and retrieval of video content within vast multimedia archives (see Figure 1.1) [ARS+18]. Additionally, content-based video analysis algorithms can provide a fine-grained analysis as well as create and find abstract relations between multiple video sources within a video database [MMK+19]. In contrast, the manual video labelling and finding of such associations in large archives are not only time-consuming for film archivists but sometimes infeasible [MMK+19]. Therefore, to allow efficient exploitation of the current historical film archives and collections, it is essential to develop tools for content-based access, search and retrieval of their films.

In the last decade, various standardisation and digitisation projects have been set in motion to encourage and support research in this area [Zec15] [IZ16]. Projects like

[1] https://www.archives.gov/ - last accessed: 31.08.2020
[2] https://www.filmmuseum.at/en - last accessed: 31.08.2020

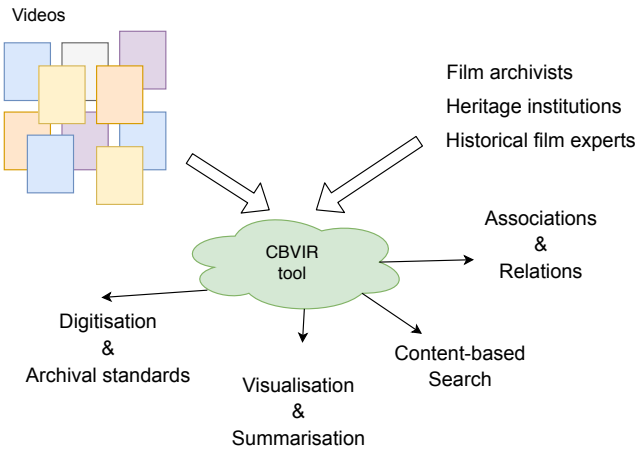


Figure 1.1: A symbolic representation of a CBVIR search system.

Ephemeral Films: National Socialism in Austria [Zec15] and I-Media-Cities [IZ16] have been carried out to ensure the preservation, accessibility and content-based searching in historical films. The projects digitise collections of ephemeral historical footage of the Holocaust and the National Socialism in Austria [Zec15] [IZ16]. However, a large number of Holocaust-related films are located in different museums and archives, with each archive supporting particular digitisation and archival standard. The lack of international archival standards represents a challenge for film archivists and historical film experts and significantly hinders the development of CBVIR tools [HK19b]. Moreover, processing the substantial amount of new, unseen film records related to the discovery and liberation of Nazi concentration camps at the end of World War II is another critical challenge. The ongoing H2020 Visual History of the Holocaust (VHH) project aims to aid and support film archivists by producing a set of international digitisation and archival standards for historical film preservation as well as tools for automatic content-based video search and analysis [CC19]. The VHH project is a joint effort of heritage institutions, historical film experts and developers to unite, standardise and illuminate existing as well as newfound film record evidence of the Holocaust [CC19]. Ultimately, the goal of the project is to supplement current knowledge with new visual records and representations of the Holocaust and its related events in that historical period [CC19].

1.2 Problem Statement

The development of an automated CBVIR tool is a challenging task considering the broad variety of video types and special video effects [ARS+18]. Moreover, developing a CBVIR tool for historical film archival adds to the complexity of the task [ZMZB11]. This is due to the quality and unique age-related properties of historical films. Archival film material is very delicate, yet many historical films are neither stored in optimal

conditions nor handled with care. Consequently, historical films often come with damaged reels, scratched and blurred frames as well as mould and contain flicker, shaking and splices [ZMB08] [ZMZB11].

The colour space of the archival films represents an additional challenge for content-based analysis [ZMB08] [ZMZB11]. Most of the historical films are grayscale and suffer from serious colour fading and low contrast from repeated replication [ZMB08]. Furthermore, the process of digitisation of archival films can also be a part of the cause for the quality deterioration of the films [ZMZB11]. With the evolution of technology, digitisation and archival of film records have become inexpensive and straightforward [ARS+18]. However, the large availability of hardware and software tools for digitisation and the lack of corresponding archival and digitisation standards significantly contribute to the complexity of the automated indexing and retrieval tool for film archives.

The majority of historical films are currently stored in museums, film archives and private collections [SZMB11]. However, there exists an abundance of film records which are neither seen nor processed by any expert. While manual video labelling is possible, it is a very time-consuming, cost and labour-intensive process for historians, film archivists and heritage institutions [HK19b]. This process also includes the chronological sorting and organisation of film collections which often contain records from several decades. Lastly, the content of historical films is diverse and can vary from ephemeral home films to the mass murder in the Holocaust. In the latter case, the film records can be a heavy burden for archivists and historical experts [HK19b]. Overall, this makes the development of an automated CBVIR tool substantially beneficial for film archivists and of crucial importance for historical film preservation and archival.

1.3 Research Questions

The primary objective of this master thesis is to bring us one step closer towards the goal of standardised historical film archival, preservation and automatic content-based video analysis and search tools. The first and foremost step in automatic content-based video analysis and CBVIR tools is the segmentation of a video into its core units i.e. shots [AM14] [Gyg18]. In the literature, this is known as the Shot Boundary Detection (SBD) problem and represents the scope of this work. The temporal separation of a video into shots is essential for CBVIR applications as it ensures fast content access to video data and prepares the data for further high-level analysis and processing [ARS+18].

The main focus of this work lies in automating the process of predicting the start and endpoints of each shot in historical videos. A shot is defined as a sequence of interrelated consecutive frames recorded by a single camera action [AM14]. Shots are combined to produce a video. The transitions between shots are classified as either Abrupt Transitions (AT) or Gradual Transitions (GT) [BGC15b] [Gyg18]. A variety of SBD methods have been proposed in the literature. The earliest SBD approaches are based on traditional computer vision methods and extract low-level frame content information such as histograms, edges and pixels (see [ZKS93] [ZMM95] [BR96]). On the

other hand, the latest SBD research shows promising results by utilizing Deep Learning and Convolutional Neural Networks (CNNs) as the base algorithm(see [SL20] [WZJ+19].

The applicability of the task of SBD is not limited to a specific domain (see [SOD10] [BGC15b] [TFK+18]). As a result, the video data used for the testing and validation of established SBD approaches stems from versatile video genres such as movies, documentaries, sports and news clips. However, only a minority of the SBD research studies dedicate their attention to historical footage [SZMB11] [ZMZB11]. One of the reasons for this could be the fact that SBD demonstrates a higher level of complexity and faces specific challenges and problems for historical films [SZMB11]. Historical films are problematic for SBD methods as they include various special properties such as scratches, flicker, instability and blur which cause a large number of false predictions [ZMB08] [ZMZB11]. All of these properties of historical films greatly contribute and have to be considered when dealing with the task of SBD. To overcome these challenges and improve the overall SBD performance in historical films, this thesis aims to answer the following research questions:

1. *How do the CNN architectural properties and feature extraction strategies affect the SBD efficiency in historical films?*
2. *How does the SBD performance in historical films depend on the training data?*
3. *What are the main challenges of historical videos concerning the problem of SBD?*
4. *How efficient are the state-of-the-art SBD approaches on historical data?*

1.4 Contributions

The main aim of this thesis is to propose and implement a novel SBD framework specifically designed for detecting the shot boundaries in historical footage. The framework is inspired by state-of-the-art SBD approaches and utilizes deep CNNs. The CNN models are trained on self-designed historical datasets. The framework follows a common three-stage approach which fundamentally consists of feature extraction, distance computation and transition (i.e. non-transition) classification [YWX+07]. To answer the first and second research question this work includes multiple training and inference experiments which investigate the effect of the training dataset, feature extraction and threshold strategy on the SBD detection performance. Furthermore, this work exploits and analyses the advantages of different architectural properties of CNNs with respect to the problem of SBD. For this reason, two types of feature extractors (a ResNet-based [HZRS16] and a VGG-based [SZ15]) are utilized for each experiment.

The performance of the SBD framework is evaluated on the historical films published in the projects Ephemeral Films: National Socialism in Austria [Zec15] and I-Media-Cities [IZ16]. This master thesis demonstrates the challenges of historical films and addresses the third research question by providing a thorough qualitative and quantitative analysis of the

attributes of historical video data and the false predictions of the SBD framework. In addition to the historical films, the efficiency of the SBD framework is validated against publicly available datasets of contemporary film material. Finally, this work includes an evaluation of the performance of the existing state-of-the-art SBD approaches on historical films. For this purpose, both traditional computer vision techniques and state-of-the-art deep learning-based approaches are taken into consideration. A comprehensive analysis and comparison of the performance of the state-of-the-art approaches and the novel SBD framework is provided and serves as an answer to the fourth research question. To summarise, the main contributions of this thesis are:

- Analysis of the historical films and creation of a historical train dataset.
- Analysis of the performance of six established SBD approaches (2 traditional computer vision and 4 state-of-the-art deep learning-based) on historical data.
- Development of a novel framework for SBD in historical videos.
- Evaluation of the proposed SBD framework on the historical and benchmark data and comparison to the state-of-the-art approaches.

1.5 Structure of the Work

The rest of the work is organized as follows. First, an introduction of SBD is provided along with a brief description of its history and its base concepts and terms in Chapter 2. Chapter 2 also reviews the state-of-the-art achievements in the field of SBD and presents the related work with a focus on SBD in historical footage. Next, the characteristics, artefacts and challenges of the historical films are discussed and presented in Chapter 3. The details regarding the methodology used in designing and developing the novel SBD framework including the descriptions of datasets used for training and evaluation are presented in Chapter 4. Furthermore, this chapter contains the implementation specifics and provides detailed information about the methods for both abrupt and gradual transition detection including the training hyperparameter configuration used. The evaluation setup and metrics, as well as the results achieved by the proposed framework and the state-of-the-art methods are compared, evaluated, and thoroughly analyzed in Chapter 5. Finally, this thesis is closed with a summary and a conclusion that highlights all achievements accomplished and illuminates the future work to come.

CHAPTER 2

State-of-the-Art

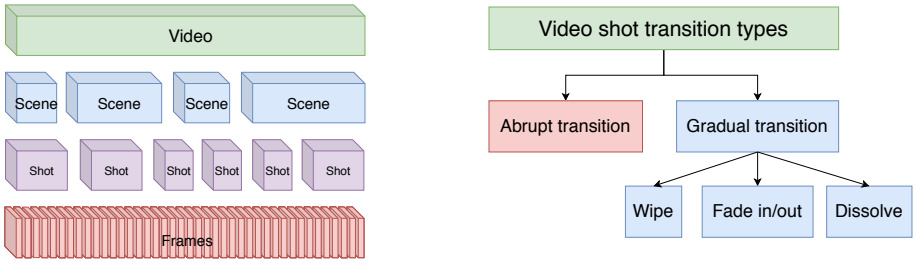
In computer science, automated video indexing and content retrieval represent an important active research area [LS13]. The problem of SBD has been a target for detailed investigation as it establishes the basis for high-level content-based video retrieval and analysis [ALBK09] [AM14] [Gyg18]. A large number of research studies have focused their attention on providing an SBD solution [ARS+18]. Consequently, many different algorithms and methods have been developed with the same goal: segmentation of a video into shots [ALBK09] [Gyg18]. This chapter provides the necessary prerequisites and describes the essential SBD related concepts, terms and definitions in detail. Furthermore, it presents the most prominent traditional and state-of-the-art SBD approaches. Finally, the chapter ends with a review of the most notable SBD contributions with a focus on historical films.

2.1 Fundamentals

Shots are elementary units of a professionally produced film [LS13]. Formally defined as a sequence of frames captured by a single device in a single continuous action in time and space, shots are regarded as fundamental low-level syntactic building blocks of a video sequence [BR96] [GKS00]. Shots are bounded by transitions, which exist between the shots [BR96]. The boundaries determine the start and the endpoint of the shot. Hence, the task of segmenting a video into its core units involves the detection of shot boundaries [YWX+07].

2.1.1 Video Hierarchy

Figure 2.1a depicts a basic hierarchical structure of a video sequence. The bottom of the hierarchy is represented by individual but temporarily ordered video frames. Video frames are the building blocks of shots [GKS00]. A group of logically connected shots



(a) The basic hierarchy of a video sequence. (b) Categorization of Transition Types.

Figure 2.1: Shots are the basic units of (semi-) professional videos. The transitions between the shots can either be Abrupt or Gradual. There exist three types of Gradual Transitions: Wipe, Fade In/Out and Dissolve.

form a scene [BR96] [VW02]. Scenes are placed one level higher in the video hierarchy as depicted by Figure 2.1a. Every scene is composed of one or more shots, which are separated by their shot boundaries. The shots within a scene can be filmed from different angles, but together the shots constitute a single semantic unit [BR96]. Finally, when combined, one or more high-level semantic scenes form a complete video sequence which is represented at the top of the video hierarchy [VW02].

For humans, the segmentation of a video into scenes is more intuitive than the segmentation of a video into shots. This is due to the fact that scenes build semantic units as opposed to the syntactic units that shots represent [BGC15a]. Furthermore, the task of segmenting a video into scenes is more complex as it requires a deep semantic understanding of the underlying video context [BGC15a]. Even though it may seem that the necessity for scene boundary detection is greater, the task of SBD is of vital importance as it separates the video into its primitive syntactic units [VW02]. These syntactic units then represent the basis for other higher-level tasks such as scene segmentation and keyframe extraction [GKS00].

2.1.2 Transition Types

To produce a video, shots and transitions are concatenated together during the video editing process [KGU10a]. Given the current advances in technology, the video editing processes allow for the generation of numerous transition types. The simplest form of transition classification divides transitions into two groups: Abrupt Transition (AT) and Gradual Transition (GT) [BR96] [GKS00] [BCS+05].

Abrupt Transitions. ATs are transitions of length one [BR96]. These transitions utilize no special effect and appear as a result of the direct concatenation of two successive shots. Hence, the transition occurs between the last frame of the previous shot and the first frame of the following shot [BR96]. ATs are known to represent sudden (sharp) changes in temporal video information and are the most frequent type of transitions.

(a) Frames 289 - 291 belong to the first shot. The next shot begins from frame 292. The AT occurs between frames 291 - 292.

(b) A sudden change in video content is visible between frame 347 and 348. Frame 347 is the end boundary of the first shot and frame 348 is the start boundary of the next shot.

Figure 2.2: ATs have no artificial effects and demonstrate a sudden change in video information.

Figure 2.2 shows an example of AT. Figure 2.2 shows a graphic representation of two ATs in historical videos.

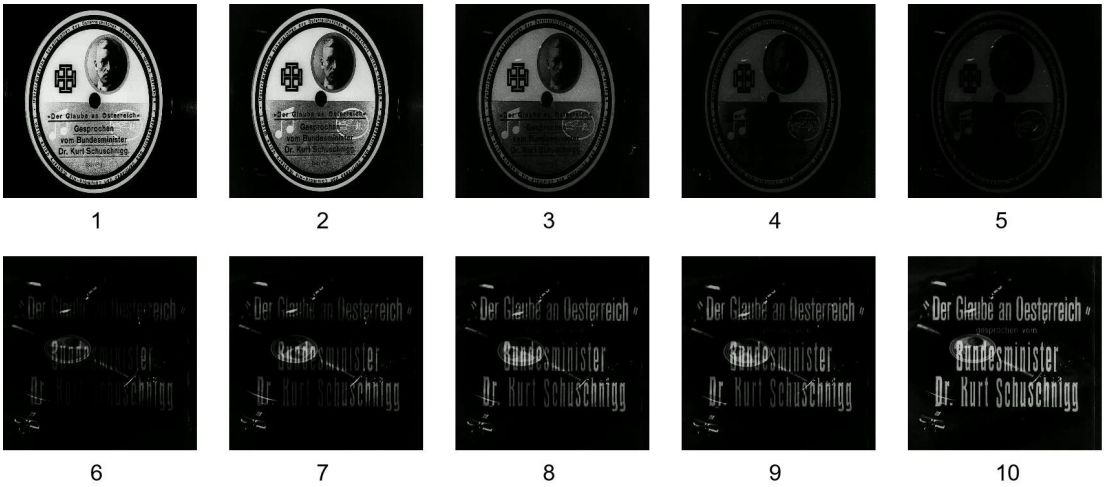
Gradual Transitions. In contrast to ATs, GTs are transitions of length greater than one. Commonly found in movies, GTs occur as a result of the utilization of special effects during the concatenation of two successive shots [GKS00]. They usually stretch over a couple of frames and depict a smooth change and contain interrelated information from the preceding and the following shot [XSX16]. Due to these reasons, GTs are more complex than ATs. Furthermore, this also impacts the detection of GTs which is a far more challenging and demanding task than the detection of ATs. Depending on the type of the effect that has been utilized to create the transition, GTs are further divided into fade transitions, wipe transitions and dissolve transitions [BR96] [GKS00] [BCS+05]. The full categorization tree of different transition types is shown in Figure 2.1b.

Fade. A Fade is a type of GT which depicts a gradual change of brightness s.t. the pixel intensities either start or end in a fixed-intensity frame (usually dark) such as a black frame [ZMM95] [BR96]. In the literature, it is distinguished between types of Fade transitions: Fade In transitions, Fade Out transitions and Fade Out-In transitions (see Figure 2.3) [GKS00].

Fade In. A Fade In (FI) starts with a monochromatic frame [ZMM95]. This type of transition takes place when the pixel intensities i.e. scene gradually emerge from the fixed-intensity frame [ZMM95]. Hence, a FI contains the frames of the next shot.

(a) Fade In: The transition starts with a monochromatic frame. Afterwards, the pixel intensities gradually emerge.

(b) Fade Out: The pixel intensities slowly change to a monochromatic frame.



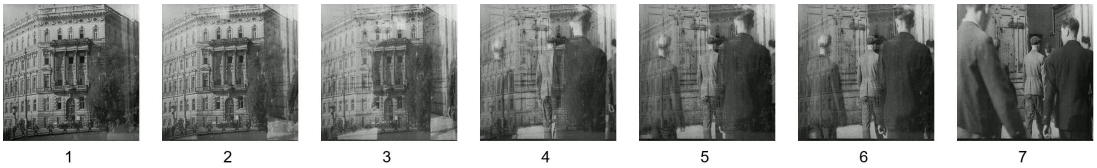
(c) Fade Out-In: A combination of a Fade Out (frames 1 - 5) after which a Fade In (frames 6 - 10) occurs.

Figure 2.3: There exist three types of Fade transitions: (a) Fade In, (b) Fade Out and (c) Fade Out-In.

Traditionally, FI transitions are used at the beginning of a movie or an act. Figure 2.3a represents an example of the gradual changes that occur in a FI transition.

Fade Out. In Fade Out (FO) the pixel intensities gradually changed from one shot and turn into a single monochromatic frame [ZMM95]. The next shot starts after the fixed-pixel intensity frame, as depicted by Figure 2.3b. Therefore, a FO transition contains only information from the previous shot. In contrast, to FI transitions, the FO transitions are typically used at the end of a movie or an act.

Fade Out-In. A Fade Out-In (FOI) is a combination of the two fade transition types [GKS00].



(a) Dissolve transition: The first shot gradually disappears, while the next gradually appears. The transition frames 3 - 6 contain overlapping content from both shots.

(b) Wipe transition: The transition frames have no content overlap and remain spatially separated. In frames 16139 - 16143, the next shot gradually replaces the first shot by pushing it off the screen.

Figure 2.4: Visual representation of the special effects of (a) Dissolve transitions and (b) Wipe transitions.

An example of a FOI transition is shown in Figure 2.3c. A FOI transition first starts with a FO. Hence, from a shot, the pixel intensities gradually change into a single monochromatic frame. From the monochromatic frame, a FI occurs in which the pixel intensities gradually emerge. The FOI is the most complex out of the three fade transition types as it contains information from the two shots i.e. the end frames of the previous shot and the start frames of the next shot. Fade out-ins are commonly used by film directors to indicate a change of scenery or time passage [GKS00].

Dissolve. A dissolve transition happens when one shot is gradually replaced by another shot [GKS00]. In other words, one shot disappears, while the next one appears. Hence, the pixel intensities of the first shot gradually diminish, while the pixel intensities of the next shot gradually increase and come into view [ZMM95]. This can be seen in the example of a dissolve transition presented in Figure 2.4a. During a dissolve transition there exist a few frames where both shots overlap [LZ01]. These frames contain parts of both shots which are interconnected and shown at the same time by increasing and decreasing the pixel intensities respectively. This makes the dissolve transitions complex and challenging to detect. Some literature refers to the dissolve transitions as a special case of the fade transitions [ZMM95]. This is due to the fact that the dissolve transitions behave like the fade transitions, except, instead of a fixed-intensity monochrome frame, the dissolve transitions involve two shots.

Wipe. A wipe transition is very dynamic and occurs when a shot replaces another one in a regular pattern such as a line [ZMM95] [BR96]. Interestingly, in wipe transitions, the two involved shots remain spatially separated at all times yet they overlap temporarily. The complexity of the wipe transitions comes from the fact that wipe transitions can

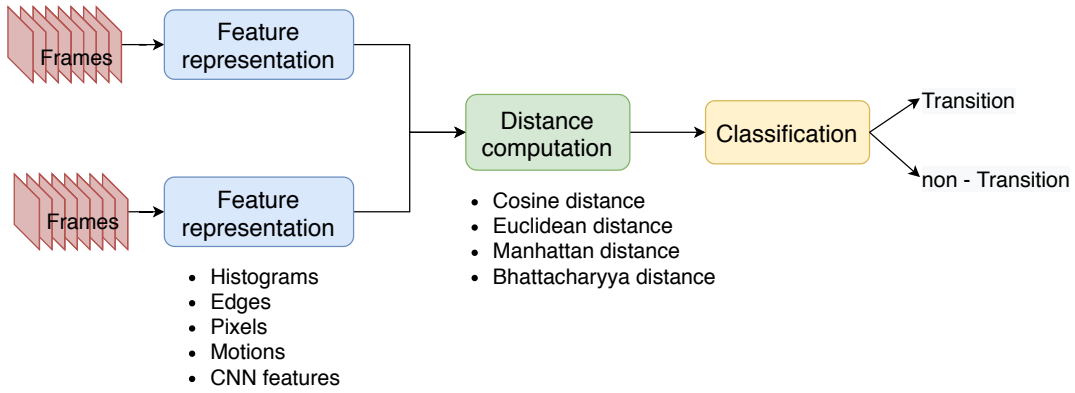


Figure 2.5: The formal SBD framework pipeline includes three main steps: representation of visual information, distance computation and transition classification.

come in numerous variations. There exist horizontal, vertical even oblique wipe transition. An example of a horizontal wipe transition is shown in Figure 2.4b.

2.1.3 Formal SBD Framework

Regardless of the detection technique utilized, at its core, each SBD method is comprised of three essential stages: representation of visual content information, construction and calculation of a (dis-) similarity measure and classification of the calculated (dis-) similarity measure value [YWX+07]. The formal SBD framework pipeline, together with the core stages are depicted in Figure 2.5. The SBD methods are, however not limited to these stages. Some of them include extensive preprocessing and/or postprocessing steps such as initial filtering and candidate selection stages (see [TFK+18] [HES+17]). This kind of additional processing is known to reduce computation costs and improve the overall detection performance of a method [TFK+18].

Representation of Visual Information

A favoured approach for representing the visual content of each video frame includes the extraction and compact presentation of frame features [YWX+07]. The major goal of this stage includes the establishment of an extraction and content representation method that is both invariant and sensitive at the same time. An invariant feature in the context of SBD is a feature that does not respond to temporal variations within a shot such as camera and object movements [YWX+07]. On the other hand, a feature is sensitive if it can capture small content details and detect the changes between the shots [YWX+07]. Ideally, a combination and the right balance between invariant and sensitive features are required to obtain an SBD method with high transition detection accuracy. However, the trade-off between these two requirements must also be taken into consideration. In the literature, there exist many different ways of visual content representation. The most

common ones include histograms, edges, pixels, motions and high-level deep convolutional features (see [GKS00] [ZMM95] [ZKS93] [TFK+18]).

Construction and Calculation of a (Dis-)Similarity Measure

The construction and calculation of a (dis-)similarity measure is the second key stage in the general SBD framework [YWX+07]. As a result, it depends and utilizes the extracted visual content from the first stage and acts as an intermediary between the first and the third stage. The conventional approach involves calculating the distance (dissimilarity/similarity) between the extracted features of two adjacent frames [YWX+07]. In the best-case scenario, the similarity measure has high values for frames within the same shot and noticeably low values for frames of shot transitions (i.e. frames which do not belong to a shot). The opposite is true for a dissimilarity measure. However, there exist many challenging disturbances such as flashes, illumination and object/camera motion that negatively affect the consistency and stability of the measure. The measures that are typically utilized in this stage include Euclidean distance, Cosine distance, Manhattan distance and Bhattacharyya distance.

Classification of a (Dis-)Similarity Value

After the calculation of a (dis-)similarity value in the second stage, the third stage performs the detection of transitions between shots. The classification of a transition against a non-transition is done according to the (dis-)similarity value that was calculated prior to this stage. When it comes to the classification process, there exist two strategies - the first strategy utilizes a fixed threshold value whereas the second strategy relies on an adaptive threshold [BCS+05]. While the first strategy requires very little effort, the utilization of a fixed threshold results in a nonrobust, nongeneralizable SBD method which will fail to perform well when applied to videos of a different video genre. To overcome this serious limitation, approaches that utilize adaptive thresholding and approaches that avoid the need for a threshold altogether are preferred [BCS+05].

2.1.4 Major Limitations and Challenges

The major limitations to the above-described framework as well as the current SBD methods include the detection of GTs, sudden changes in illumination such as flashes and camera and object movements [BCS+05] [YWX+07]. To obtain a well-performing and accurate SBD method, these challenges must be addressed. Nonetheless, the true complexity of the task of SBD lies in the successful overcoming of these challenging issues.

Sudden Illumination Changes. Many of the popular extraction and representation of visual content methods are based on colour features [YWX+07]. Luminance is an integral characteristic of colour which depicts the amount of light intensity as perceived by the human eye [OR06]. This makes SBD methods which rely on colour-based features very sensitive to illumination changes. As a result, these SBD methods often mistake

sudden illumination changes for AT. To counteract this issue, several illumination-invariant features have been proposed [QWG03]. Even though illumination-invariant features succeed in dealing with a great number of illumination issues, SBD methods lose a significant amount of luminance information necessary for transition detection [YWX+07].

Camera and Object Movements. Apart from transitions, which are placed and exist between the shots of a video sequence, significant alteration of visual content information can be caused by (large) camera and object movements [BCS+05] [YWX+07]. While such movements are rarely confused for an AT, the opposite holds when it comes to GT. Continuous slow-motion mimics the similarity patterns of an artificial effect. Due to the similar amount of content variation, SBD methods have trouble distinguishing between an actual GT and these kinds of movement. One possibility of dealing with this challenge is the integration of features for motion compensation[YWX+07].

GT Detection. The formal SBD framework works well for AT detection, however, it comes short when it comes to GT detection [YWX+07]. As described in Section 2.1.2, GTs include a number of different special effects. Depending on the type of (dis-)similarity measure used, different types of effects admit different value trends and patterns[YWX+07]. Furthermore, the duration of a GT does not follow any scheme and is not bound by the number of frames. Hence, there is no way of generalizing or predicting the duration of a certain GT. Although the frames of GT carry lower similarity values than the frames within a shot, these values are not as drastically low as the ones of ATs. Lastly, the similarity values of GT frames show similar patterns to similarity values between frames which include camera and object movements [BCS+05]. All of these reasons make the detection of GT extremely challenging and complex.

2.2 Traditional SBD Approaches

Most of the current state-of-the-art SBD methods take advantage of the success in the field of Deep Learning and utilize CNNs as a base algorithm (see [Gyg18] [TFK+18] [HES+17]). This greatly differs from the original SBD techniques which rely on conventional computer vision methods (see [BR96]). Consequently, the classification of these approaches depends on the type of visual information extracted. Typically popular approaches utilize pixels [ZKS93], edges [ZMM95] [ALBK09] and histograms [LLZ16] [KGU10b] as a representation of the frame content. Further approaches include the utilization of transform-based features [CFAC03] [UGE06] as well as the extraction of motion-based [ZKS93] [BGG99] and statistical-based [Han02] features from the video frames.

Pixel-based methods are one of the earliest methods for detecting the shot boundaries in videos [ZKS93]. They deal with the problem of SBD by extracting and calculating the difference between the pixel intensities of consecutive video frames [ZKS93]. To declare a shot boundary, the sum of absolute differences between the pixel intensities is calculated and compared against a threshold. Overall, pixel-based techniques are slow and very sensitive to camera and object movement [BR96].

On the other hand, approaches based on histograms are particularly popular and widely used (see [LLZ16], [KGU10b], [MF03]). The base method creates a grayscale or colour histogram for each frame and computes the bin-wise distance [MF03]. A transition is detected if the computed distance exceeds a threshold. Nonetheless, almost every variation of calculating the intensity of colour histogram differences has been proposed in the literature [GKS00] [MF03]. The proposals include the exploration of chi-square tests and different colour spaces such as RGB, HSV, YIQ, and etc. Histogram-based methods are not as sensitive to motion as pixel-based methods and have proven to be effective for AT detection [ZKS93].

A more recent method that is based on the multilevel difference of colour histograms is proposed by Li et al [LLZ16]. The method detects ATs as well as GTs using two different thresholds respectively. The approach is separated into three stages. The first stage extracts the colour histogram from the frames. Next, the Euclidean distance between the colour histograms of consecutive frames is calculated. To check whether a GT exists, the multilevel distance between a set of consecutive frames is computed. The calculated distance values are then filtered with two thresholds and produce candidate frames for AT and GT. Furthermore, this stage removes the frames between the start and end of a GT and thus creates a new sequence. This allows the method to treat GTs in the same way as ATs. The second stage removes noise by filtering out the local maximums. In the last stage, a voting mechanism is employed that makes the final decision if a transition occurred. The performance is evaluated on the TRECVID 2001 [SOD10] dataset on which the method achieves an F1 score of 89% [LLZ16].

Edge-based methods rely on the number and position of edges in successive video frames, for the proper detection of ATs and GTs [ZMM95]. Zabih et al. [ZMM95] introduce the concept of Edge Change Fraction which measures the proportion of entering edge pixels against the proportion of exiting edge pixels between two consecutive frames. High values of the edge change ratio indicate a transition. Zabih et al. also classify the transition type as AT, wipe, dissolve and fade by further analyzing the spatial distribution and the value of the edge change fraction. The proposed algorithm starts with Canny's algorithm [Can86] which results in two binary images. Next, Gaussian smoothing is applied to improve the efficiency of the algorithm. The entering and exiting edge pixels are counted and the value for Edge Change Ratio (ECR) is computed. The major limitations of edge-based methods are the computational cost and the sensitivity to camera operations such as zoom, pan and tilt [ZMM95].

In contrast, transform-based methods work fast by transforming the signal i.e. frame from the time (spatial) domain into the transform domain [ARS+18]. Examples of transforms include discrete cosine transform (DCT), discrete Fourier transform (DFT) and discrete wavelet transform (DWT). Transforms differ in their basic function which is responsible for extracting the features from the signal [ARS+18]. Some of the more prominent transform-based approaches include the proposals by Cooper et al. [CFAC03], Urhan et al. [UGE06], Zaharieva et al. [ZMZB11]. and Priya et al. [GGD12].

Sophisticated motion-based approaches calculate the optical flow and rely on the number

Approach	Feature	Distance Metric	Threshold	Detection	Dataset	F1-score
Zhang et al. [ZKS93]	Pixels	City Block	Fixed	AT, Dissolve	3 videos	96%
Li et al. [LLZ16]	Histogram	Euclidean	Adaptive	AT, Dissolve, FOI	TRECVID01	87%
Küçüktunç et al. [KGU10b]	Histogram	Fuzzy rules	Fixed	AT, Dissolve, FOI	50 MPEG-7 seq.	77%
Mas and Fernandez [MF03]	Histogram	City Block	Fixed	AT, Dissolve, FOI	TRECVID03	83%
Zabih et al. [ZMM95]	Edges	ECR	Fixed	AT, GT	50 MPEG-7 seq.	93%
Priya et al. [GGD12]	WHT	City Block	Fixed	AT	TRECVID	92%
Urhan et al. [UGE06]	DFT	Correlation coeff.	Adaptive	AT	114 MPEG-7 seq.	92%
Cooper et al. [CFAC03]	DCT	Cosine similarity	Fixed	AT, GT	TRECVID02+03	80%
Hanjalic [Han02]	Statistical	Mean absolute error	Fixed	AT, Dissolve	5 videos	96%

Table 2.1: Summary of the traditional SBD approaches.

of motion vectors for the detection of transitions [ZKS93] [GKS00]. Motion vectors are extracted by dividing frames into regions after which block matching algorithms are applied [GKS00]. However, motion vectors as features are unstable and alone insufficient for the successful detection of transitions [BR96]. Lastly, approaches based on statistical comparison divide the video frames into blocks [Han02]. The blocks of successive frames are then compared using statistical techniques such as mean, deviation, variances and likelihood ratio [ZKS93] [BR96] [Han02]. If a certain amount of blocks exceed a given threshold value, a transition is declared. Due to the computational complexity, these type of approaches are slow and produce a high number of false detections. The traditional SBD approaches are summarized in Table 2.1

2.3 Deep Learning-based Approaches

One of the first CNN-based SBD methods is developed by Xu et al [XSX16]. The proposed approach detects ATs and GTs and follows the three-stage framework. It includes a candidate segment selection stage, a CNN-based feature extraction stage and a novel transition classification stage. The candidate segment selection stage is performed using adaptive thresholding. Furthermore, Xu et al. also utilize an AlexNet inspired CNN model in the feature extraction stage. The distance between two frame feature vectors is calculated with the cosine distance. The last stage involves a threshold-dependent classification of both ATs and GTs. The approach is evaluated on the TRECVID01 [SOD10] dataset and achieves 98% F1-score on the AT detection and 96% F1-score on the GT detection task [XSX16].

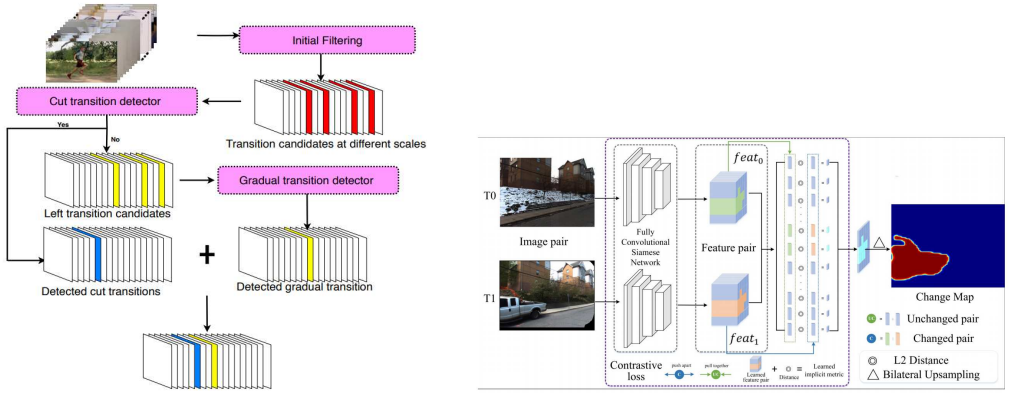
Baraldi et al. [BGC15a] propose an SBD framework in which they utilize both visual and textual frame features. The authors rely on a Siamese network for the visual feature extraction. The extraction of the textual features is performed using Skip-gram [MSC+13] models which represent words into feature vectors. The distance between the features is calculated using Cosine similarity. In the end, both the visual and textual features are used to construct the final similarity score. The similarity scores computed by the network are combined to create a similarity matrix. The matrix then serves as input to a spectral clustering algorithm which outputs the final shot boundaries. The framework is evaluated on the BBC Planet Earth dataset and achieves an F1-score of 62% [BGC15a].

Hassanien et al. [HES+17] first propose a deep neural network which trained end-to-end and call it DeepSBD. The network utilizes C3D [TBF+15] convolutions and can classify a segment of frames into three classes: sharp, gradual and no transition. The input of the network is a 16-frame segment with 8 frames overlap. The network consists of five 3D convolutional layers, and its output is sent to an SVM classifier to classify the CNN features into the three classes. Afterwards, segments with the same label are merged and passed to the post-processing stage, where a colour histogram and Bhattacharyya distance are computed to assist the decision making. The network is trained on two datasets, one synthetically generated and one with complex hard-negative cases. DeepSBD is evaluated on the RAI [BGC15b] and TRECVID [SOD10] datasets and achieves significant improvement in detection performance and processing speed [HES+17].

Concurrently to Hassanien et al. Gygli [Gyg18] implements a C3D [TBF+15] based CNN architecture model for fast SBD. The main advantage of using a 3D convolution instead of the standard 2D convolution is that the input contains multiple frames instead of a single frame. This means that both the spatial and temporal information is preserved and examined at the same time. The proposed CNN architecture model utilizes 3D convolutional layers only, which makes it fully convolutional and compact. The network architecture consists of 4 convolutional layers, each followed by a ReLU activation function with a softmax function at the end of the network. The network was trained end-to-end with a newly created dataset that consists of millions of frames. The dataset consists of ATs such as hard cuts and crop cuts as well as GTs such as dissolves, wipes and fade transitions. It includes 79 videos with a total duration of 3.5 hours. At inference, the network analyzes 100 frames at a time with an overlap of 9 frames and classifies each frame as either same shot with the previous frame or not. Gygli evaluated the method on the RAI [BGC15b] dataset and reported precision and recall values greater than 90% on most of the videos [Gyg18].

Triggered by the success of the previous approaches Tang et al. [TFK+18] propose a different approach. Their cascade framework includes an initial filtering module and two separate targeted detectors for ATs and GTs respectively. Figure 2.6a depicts the framework together with its stages and core components. The initial filtering module utilizes the features extracted from SqueezeNet in combination with adaptive thresholding to produce a set of transition candidates. The selected candidates are forwarded to the AT detector which represents a 2D CNN model trained to learn a similarity function between two frames and detect ATs. The input of the AT detector, is an image pair concatenated together as a 6-channel image. In the last step, the remaining candidates are passed to the GT detector. The GT detector is implemented as a C3D-based [TBF+15] CNN model with ResNet-18 as a backbone of the network. Furthermore, the authors present the first large-scale dataset for SBD, called ClipShots [TFK+18] which they use for training and evaluation of the model. Apart from ClipShots, Tang et al. also evaluate their framework on the RAI [BGC15b] and TRECVID07 [SOD10] datasets on both of which they achieve outstanding results.

Guo et al. [GFZ+18] propose a deep neural network framework for Scene Change Detec-



(a) Framework with initial filtering and an AT and GT detector [TFK+18]

(b) Architecture of CosimNet [GFZ+18]

Figure 2.6: Deep learning-based detection pipelines for: (a) SBD using two targeted detectors by [TFK+18] and (b) SCD using a Siamese network by [GFZ+18].

Approach	CNN Architecture	Input	Detection	Dataset	F1-score
Xu et al. [XSX16]	AlexNet inspired	Segments of 6 frames	AT, GT	TRECVID01	99% AT, 97% GT
Baraldi et al. [BGC15b]	Siamese Network	Image Pair	AT, GT	RAI	84%
Hassanien et al. [HES+17]	C3D-based network	16 frames w. 8 frames overlap	AT, GT	RAI	94%
Gygli [Gyg18]	C3D-based network	100 frames	AT, Dissolve, FOI	RAI	88%
Tang et al. [TFK+18]	Image concatination (AT) + C3D (GT)	6 images concat. + frame segments	AT, GT	RAI	93.5%
Guo et al. [GFZ+18]	Siamese Network	Image Pair	AT	CDNet Scene Dat.	85%

Table 2.2: Overview of the state-of-the-art SBD approaches.

tion (SCD). This means for a given input pair of images, the framework outputs a change map highlighting the detected changes in the scene. Since the change map represents the change between the input image pair, this enables the framework, the ability to detect ATs in videos. Thus, it can be applied to the problem of SBD. The authors address several critical SCD challenges such as illumination, shadows and camera viewpoint differences by developing a fully convolutional siamese network called CosimNet [GFZ+18]. The proposed network can directly compare and detect the differences in an image pair by extracting the convolutional features and calculating the distance between them. The architecture of CosimNet is depicted in Figure 2.6b. To improve the overall performance, the authors integrate the learning of an implicit distance metric. The main idea of the implicit distance metric is to push apart changed pairs and pull together unchanged pairs. To achieve this, the authors define and utilize contrastive loss. The final output of the network is represented by the change map that visualizes the changes between the input image pair.

An overview of the comparison points of the presented state-of-the-art SBD methods is provided in Table 2.2.

2.4 SBD in Historical Films

Regardless of the yearlong extensive SBD research, there is an evident lack of studies that dedicate their resources into developing SBD methods and techniques that support and mitigate the effects and challenges of mediocre grayscale films [SZMB11]. Furthermore, even less research has been devoted to investigating SBD in historical film material specifically. Historical film material carries a unique set of artistic and technology-related properties that pose challenges to existing SBD methods [ZMB08]. However, no large-scale qualitative evaluation and analysis of established SBD methods on historical film material has been observed in the literature. Finally, only a small effort has been made to prevent the challenges of historical films from affecting the overall SBD performance [UGE06] [ZMZB11] [SZMB11].

Urhan et al. [UGE06] investigate the problem of SBD in archive films. The authors propose a novel phase correlation-based SBD framework optimized for AT detection in archive material. The framework consists of three steps. The first step involves spatial subsampling of the video frames after which the phase correlation between consecutive frames is calculated. In the second step, ATs are classified using a double threshold strategy which involves a global as well as local adaptive threshold. In the last step, false detections are removed by the means of a heuristic procedure which relies on mean and variance tests. The effectiveness of the framework is evaluated on mainly grayscale archive films which stem from the beginning of the 20th century. Even though the films contain many visual defects and degradations, the framework achieves 99% recall and 98% precision on the archive film material [UGE06].

Zeppelzauer et al. [ZMB08] examine the historical films by the famous Soviet filmmaker Dziga Vertov. The films are documentaries from the 1920s and contain political, social and economic events. The majority of the films are silent grayscale films which run at 16 frames per second. Through the films, Zeppelzauer et al. analyze the properties and artefacts characteristic for historical film material. Furthermore, the authors investigate the importance and effects of such artefacts on the process of automated video analysis. The authors point out that the historical films by Vertov commonly include various spatial effects as well as multi-image compositions to attract and engage the attention of the audience [ZMB08]. Additionally, the authors note the historical films contain a large number of complex GTs. Lastly, Zeppelzauer et al. also investigate and develop techniques to support the process of automated film restoration [ZMB08].

Zaharieva et al. [ZMZB11] study the problem of SBD in historical film material and focus on the challenges historical films pose for the problem of SBD. The authors investigate the effect of preprocessing as a way to overcome the difficulties which are caused as a result of the poor video quality and the artefacts historical videos contain. However, in the paper, they conclude that the frame preprocessing step either leads to information loss or introduces additional noise to the frames. Moreover, Zaharieva et al. develop a technique for the detection of intertitle frames. The technique utilizes the edge and intensity histograms which are extracted from each frame. These are then classified by

a linear SVM classifier. The proposed method for intertitle frame detection achieves 97% accuracy. Additionally, the authors propose a method for the detection of black frames which performs with 93% accuracy. For the problem of SBD, they propose a novel AT detector based on a DCT feature and an edge descriptor. Lastly, the SBD detector is evaluated on 8 hours of historical films on which it achieves over 0.91% F1-score [ZMZB11].

A method for the detection of GTs in historical film material is proposed by Seidl et al [SZMB11]. In their paper, Seidl et al. address the challenges which occur when GT detection is performed on historical films. To provide complete support for grayscale historical films, in their study, the authors experiment with different feature combinations, distance measures and classification models. The proposed method follows the common 3 stage detection framework but also includes an additional fourth stage called post-processing. In the first stage of the SBD framework, the visual content of each frame is extracted in terms of luminance and edge histogram. The second stage calculates the distance between the features for which Euclidean distance, Chi-Square distance and Cosine similarity are employed. To boost the performance of the GT detection, the distance values between several features are utilized. In the third stage, an SVM classifier is used to distinguish between a GT and non-GT. Finally, the post-processing stage is used to verify the true positives and reduce the number of false positives. This is done by creating a similarity matrix that represents a combination of all pairwise similarities in a given frame range [SZMB11]. The proposed method is evaluated on the historical films produced by Vertov [Lip19] and the TRECVID [SOD10] 2006 dataset. Seidl et al. report that the approach significantly improves GT detection in historical films. Lastly, the proposed method achieves 56.2% F1-score on the benchmark dataset TRECVID06 [SOD10], although the model was not optimized for contemporary film material [SZMB11].

One of the latest papers that focuses on SBD in historical material is published by Helm and Kampel [HK19a]. The authors propose a state-of-the-art inspired SBD framework which is geared toward AT detection in historical films. The framework relies on deep learning and comprises three main stages: candidate selection, CNN-based feature extraction and similarity comparison. The first stage represents a base implementation of DeepSBD configured with a segment size of 16 frames with an 8-frame overlap. In the second stage, the authors experiment with and extract the features from 3 CNN-based feature extractors: VGG19, ResNet and SqueezeNet. The last stage measures the similarity of consecutive frames. As a similarity measure, Helm and Kampel utilize and test Euclidean Distance as well as Cosine Similarity. The performance of the framework is evaluated on a historical dataset as well as ClipShots [TFK+18]. The authors report superior performance on the historical data with VGG19 as feature extractor and Cosine Similarity and achieve a recall value of 84% and a precision value of 89% [HK19a].

Finally, Helm and Kampel [HK19b] also produce a follow-up paper in which they further investigate the historical film material. In their paper, Helm and Kampel target the problem of Shot Type Classification (STC) in historical films. The authors utilize a

Approach	Method	Detection	Data	F1-score
Urhan et al. [UGE06]	DFT + Correlation coeff.	AT	1009 archive films	99% AT, 94% GT
Zaharieva et al. [ZMZB11]	DCT + Edge Desc.	AT, Intertitle and Black Frames	Vertov Collection	91% AT, 97% Intertitle, 93% Black
Seidl et al [SZMB11]	DCT + Edge Desc.	GT	Vertov Collection	54.5%
Helm and Kampel [HK19a]	CNN Features + Cosine Similarity	AT	Ephemeral Films	86.6%
Helm and Kampel [HK19b]	CNN model	STC	Ephemeral Films	71.2%

Table 2.3: Research studies and SBD approaches with a focus on historical data.

CNN-based algorithm to classify each shot into one of the four shot type categories: Extreme-Long-Shot (ELS), Long-Shot (LS), Medium-Shot (MS) and Close-Up (CU). The method archives 70% recall and 72% precision on the historical dataset [HK19b].

Table 2.3 outlines the research studies and SBD approaches with a focus on historical film material.

CHAPTER 3

Historical Film Material

The historical film material explored in this thesis is challenging for SBD due to its quality and its age-related properties. This chapter first provides basic background information about historical films. It states the origin of the historical datasets and presents the general theme and topic of the films. Next, Section 3.2 reviews the properties and most important characteristics of historical films. This chapter also describes the artefacts found in the films. Finally, it discusses how the artefacts challenge and impede the process of SBD in Section 3.3.

3.1 Background

The historical video datasets used and investigated in this thesis consist of archive films with a focus on the Holocaust and its related events. The historical films portray the theme and ideology of National socialism and capture the history, development and lifestyle of the city of Vienna, Austria during that period [Zec15] [IZ16]. The two datasets called EFilms and IMC, contain 66 and 78 historical video films respectively. The films of the EFilms dataset were published during the project Ephemeral Films of the Austrian Film Museum [Zec15]. This dataset represents a unique collection of films including home videos, educational and advertisement films. The films are raw and document life before the Holocaust and include the rise and acts of the National Socialist German Workers' Party [Zec15]. On the other hand, the films of the IMC dataset were published as part of a European film archive initiative called I-Media-Cities [IZ16]. These films explore the visual history, urban planning, sociology and anthropology of the city of Vienna, Austria [IZ16]. In terms of chronological timeline, the films are set and take place in the years between 1910 and 1960. The shot boundaries in both of the datasets are annotated by experts and historians of the VHH project consortium [CC19].

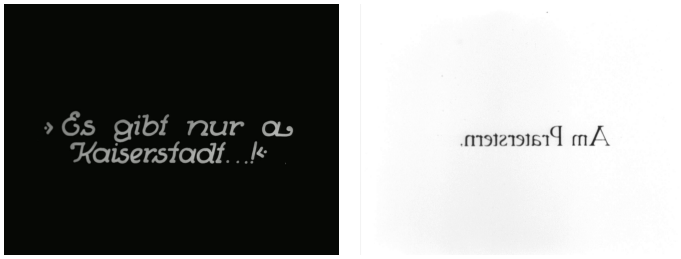
3.2 Characteristics

Since the historical films utilized in this thesis originate from and have been produced at the beginning of the 20th century, all of them are silent. Furthermore, most of the films are grayscale, only some of the newer videos incorporate colour. The historical films of both datasets are of varying length. The original historic film material of the EFilms dataset is collection of 9.5mm, 16mm and 35mm grayscale films most of which are based on easily inflammable nitrate cellulose [Zec15]. The EFilms video films have been digitized with a resolution of 960x720 pixels and a frame rate of 24 Frames Per Second (FPS) [Zec15]. In contrast, the original analogue format of the IMC films is represented by 35mm thick grayscale films [IZ16]. The analogue films of the IMC dataset have been digitized with a resolution of 1440x1080 pixels. Only a small subset of the films have a frame rate of 24 frames per second. The rest of the films are played at non-standard frame rates which vary from 16 to 22 frames per second. Further characteristics of historical film material observed in the two datasets include the utilization of intertitle frames, the common occurrence of black, grey and white frames as well as the integration of complex and uncommon gradual transition types.

3.2.1 Intertitle Frames

An important concept of historical film material and especially silent archive films is the utilization of intertitle frames [ZMZB11]. At the beginning of filmography, intertitle frames were integrated to make up for the lack of sound [ZMZB11]. Hence, intertitle frames provide the semantic context and description that the screen itself is unable to capture. In historic film material, intertitle frames are used to set the scene, time and location and introduce new characters [ZMZB11]. Essentially, intertitle frames prepare the viewer for the next scene by providing him with the necessary background information required for semantic understanding. Furthermore, intertitle frames can separate and divide one topic from another, temporally structuring the film [ZMZB11].

Intertitle frames are defined as frames with monochromatic text on a monochromatic background [ZMZB11]. In the EFilms and IMC datasets, the most common type of intertitle frames includes white text on a black background as demonstrated by Figure 3.1a. A less common type of intertitle frames includes a black text on a white background (see Figure 3.1b). This type of intertitle frames is found in some of the videos of the IMC dataset. In contrast to intertitle frames in the historical film material, intertitle frames in contemporary videos contain little text and are typically used for acknowledgements, opening and closing credits [CS02]. However, since the need for intertitle frames has significantly dropped, their appearance in modern films and videos is scarce. Nonetheless, intertitle frames did not die out completely and can still be seen in movies as a conclusion to the story and some television shows [CS02] [ZMZB11].



(a) (b)

Figure 3.1: Different types of intertitle frames.

(a) (b) (c)

Figure 3.2: Common occurrence of black, grey and white frames.

3.2.2 Black, Grey and White Frames

Black frames as the name suggest, are frames which are entirely black. Typically, black frames are employed either at the beginning or at the end of a FI i.e. a FO transition [ZMZB11]. Apart from this justified utilization, in historical video material black frames also appear quite often either to separate shots and semantical content or with an artistic goal [ZMZB11] [ZMB12]. Figure 3.2 shows examples of these cases. In the two datasets EFilms and IMC, similar cases are observed with the utilization of both grey frames and white frames.

3.2.3 Gradual Transitions

An interesting characteristic of historical film material is the utilization of complex GT [SZMB11]. In historical video material, the length of GTs is significantly greater as opposed to the GTs used in contemporary video material. Furthermore, the GTs integrated into historical films are remarkably complex and sometimes even represent a combination of two GT types. This is observed in the historical films from IMC dataset. An example of a complex GT case is illustrated in Figure 3.3.

Another peculiar feature attributed to historic films is the utilization of extraordinary types of GT which are rare and infrequent in modern films and videos [SZMB11] [ZMB12]. Examples of GT types, popular in historic films include the Iris transitions. The Iris

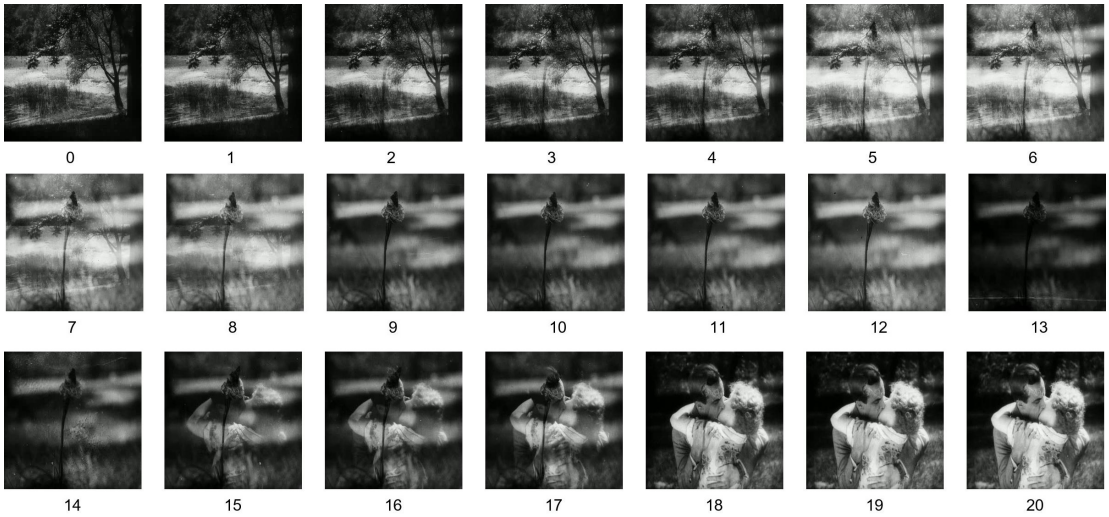


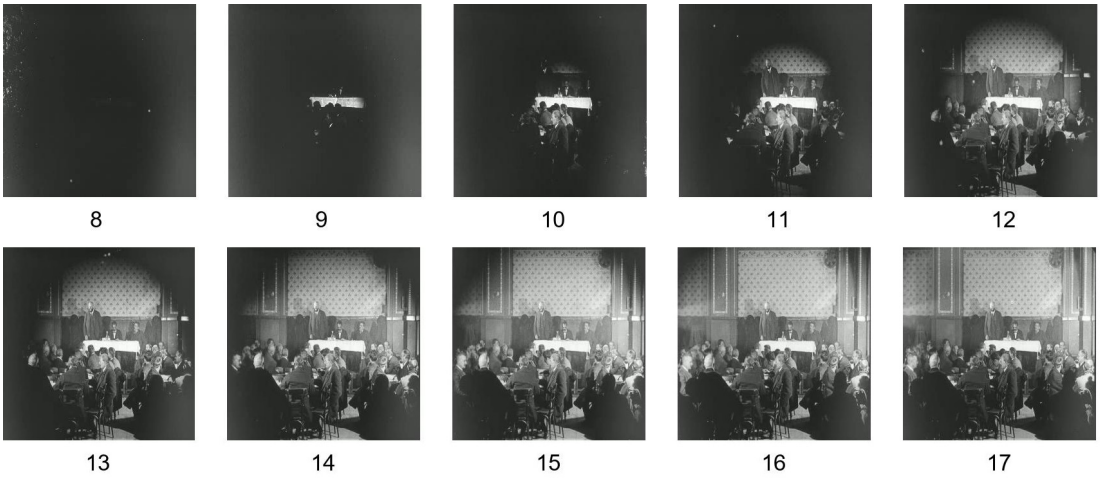
Figure 3.3: An example of a complex GT transition found the IMC dataset. The transition frames include two dissolve effects occurring sequentially.

transitions belong to the category of Wipe transitions [ZMB12]. It is distinguished between Iris In transitions and Iris Out transitions [ZMB12]. In an Iris In transition, the existing image moves into a circle that gradually decreases in diameter until it vanishes completely off the screen. At the same time, the area around the may begin to show the new shot which increases until it becomes fully visible and takes up the whole screen (see Figure 3.4a). The Iris Out transition represents the opposite process of an Iris In transition. In an Iris Out transition, the image is slowly replaced by a circle that shows the new shot. The circle then gradually increases in size until it takes up the whole screen, as shown by Figure 3.4b.

3.2.4 Special Effects

Apart from the intertitle frames and the uncommon GT, historical films are also recognized for their unique special effects [Fos18]. Stop-motion is a popular technique used often in archive films to animate objects and intertitles [ZMB12] [Fos18]. Additionally, this method was used to animate and produce hand-drawn animation films i.e. cartoons [Fos18]. Occurrences of hand-drawn animations are discovered in some of the historical films of the IMC dataset. Figure 3.5 shows sever examples of the hand-drawn animations found in the IMC dataset. Today, the craft of hand-drawn animation as well as the utilization of the stop-motion technique have been completely replaced by computer-generated animations with sophisticated 2D and 3D computer graphics [Fos18].

(a) Iris In: The frame content moves into a circle which gradually decreases. The transition ends with a monochromatic frame.



(b) Iris Out: A monochromatic frame is replaced by a circle which shows the content of the new shot. The diameter of the circle slowly increases until it takes up the whole screen.

Figure 3.4: Iris transitions belong to the group of Wipe transitions and are very common in historical films. There exist two types of Iris transitions: (a) Iris In and (b) Iris Out transitions.

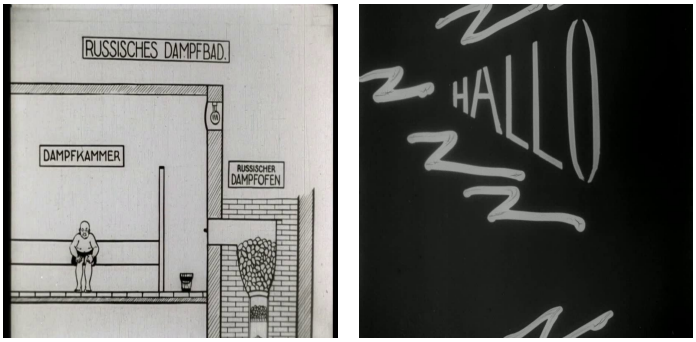


Figure 3.5: Special effects found in the datasets

3.3 Artefacts and Challenges

The original filmstrips of the historical video material used in this thesis have been produced several decades or some even a century ago [Zec15] [IZ16]. Since most of the films are ephemeral, the actual value and importance of the filmstrips were estimated only later. Therefore the filmstrips were neither stored in optimal conditions nor handled with care. The historical films used in this work are the result of multiple generation copies [Zec15] [IZ16]. Their quality is substandard as they were replayed, copied and suboptimally stored over the years.

Over the years, film archives have acquired the required expertise in preserving historical films [Fos18]. Nonetheless, the filmstrips are made of nitrate cellulose and are prone to decay. The nature of filmstrip decay can be both physical and chemical [Fos18]. The consequences, however, vary and include shrinkage, scratches colour and contrast loss, tears and flicker [ZMB08] [SZMB11] [Fos18]. To mitigate these consequences restoration and replication have become an integral part of the preservation process which to some extent amplify and contribute to the filmstrip damages. In this context, an erroneously created new image element, as well as an erroneous removal of an original image element, is referred to as artefact [Fos18].

3.3.1 Local Artefacts

One side-effect of the organic compounds found in the filmstrips, as well as incorrect storage over the years, is the shrinkage and contraction of the filmstrips along the vertical and horizontal axis [ZMB12]. This, in turn, results in frame displacements and distortions which can be noticed in both datasets EFilms and IMC. A consequence of vertical filmstrip contraction is the visibility of frame lines [ZMB12]. Furthermore, the extent of vertical shrinking in some films of the datasets is so great, that it resulted not only in visible frame lines but also in the visibility of the previous and the next frame content. Examples of the consequences of vertical shrinking are provided by Figure 3.6a and Figure 3.6b. The side effects of horizontal filmstrip shrinking, on the other hand, include the visibility of the perforation of the filmstrip when the extent of the shrinking surpasses a certain limit. This can be observed in the examples provided by Figure 3.6c. Another possible reason for the visibility of the frame lines and their perforations as well as frame distortions could be the copying of misaligned film strips [Fos18].

Filmstrips can endure great chemical damage if they are preserved at incorrect and inadequate conditions [Fos18]. A high-temperature level or increased humidity can cause the films i.e. the film emulsion to decompose [Fos18]. This can be witnessed in a small subset of the films in the two datasets. Affected frames suffer either partial or complete information loss and look as though the film emulsion has melted. A visual representation of such cases is depicted in Figure 3.7. Inadequate storage conditions can also result in mould on the filmstrips which causes image distortions (see Figure 3.7) [ZMB08] [ZMB12] [Fos18]. Furthermore, inadequate storage conditions can create the perfect environment for the

(a) (b) (c)

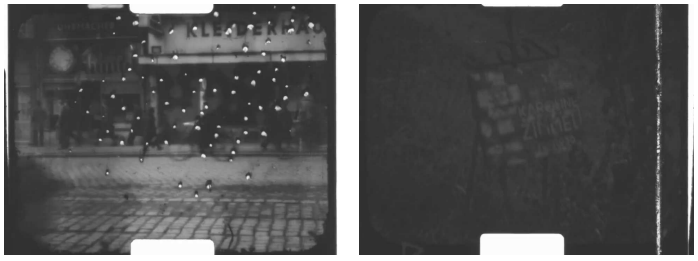
Figure 3.6: The effects of filmstrip shrinkage: visible frame lines, perforations and parts of consecutive frames.

Figure 3.7: Spills, mould, melting, bacteria found in historical films.

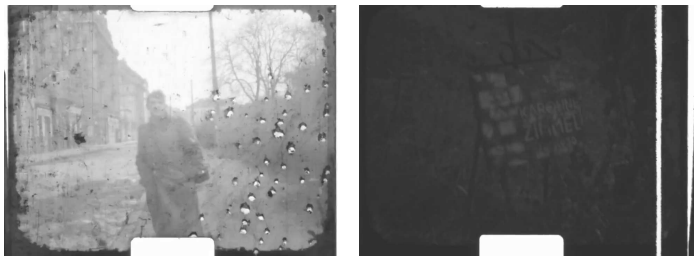
reproduction and growth of fungus and other bacteria in the film itself [Fos18]. In this case, the bacteria can grow to use the film as food, the damage of which is irreversible.

Continuous playback of the films is also known to cause artefacts [ZMB08] [ZMB12] [Fos18]. Running the films through old projectors as well as the dirt inside them can cause scratches deep enough to remove the film emulsion [Fos18]. The most common type of scratch, known as vertical scratch, covers multiple adjacent frames at the same place throughout the filmstrips. An example of this artefact is shown in Figure 3.8d. Furthermore, any other kind of physical damage caused by the projectors, such as small scratches also known as brights or the integration of dirt particles (darks) is commonly referred to as dust (see Figure 3.8c) [Fos18].

Another type of artefact that also causes information loss is the cue dot. Cue dots represent punched in holes in the film reels made by projectionists [Fos18]. Cue dots were used as a signal to start the projector at which the reel has been loaded [Fos18]. Throughout the



(a) Bright scratches



(b) Bright and Dark scratches

(c) Dark particles (d) Vertical scratch

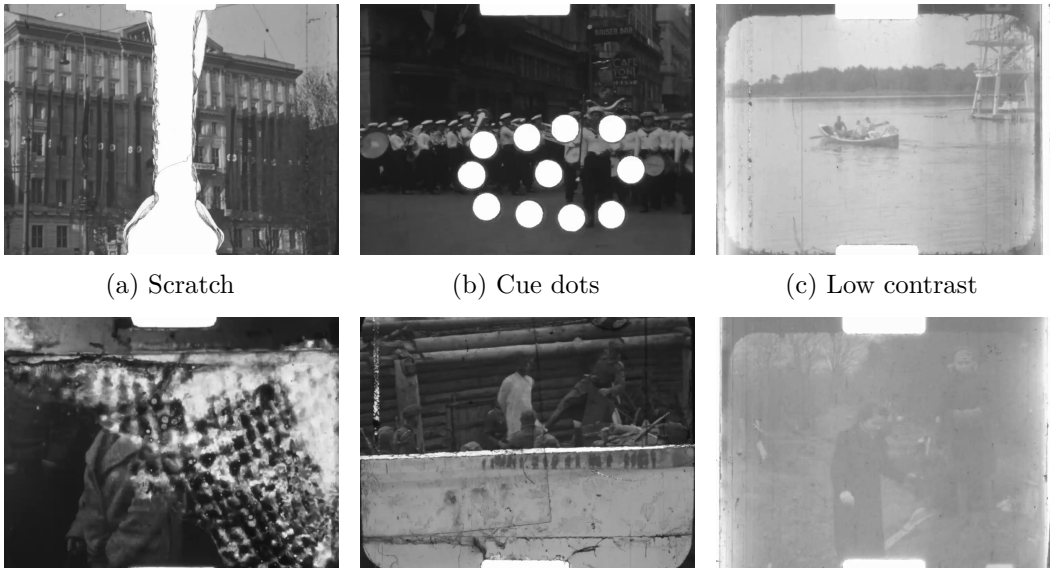
Figure 3.8: Physical artefacts of the filmstrips: scratches and dust.

EFilms and IMC datasets, many such cue dots were encountered. Figure 3.9b includes some encounters of the cue dots in the historical films. Filmstrips that have been through multiple projections by many projectionists tend to acquire many cue dots and similar perforations.

3.3.2 Temporal and Global Artefacts

Playback and presentation of the films can also be the cause of more serious damage [ZMB12] [Fos18]. The replay of the films carries the risk of filmstrip tear [ZMB12] [Fos18]. If a few frames are destroyed in the tear, the tear leads to permanent information loss. In this case, when the film is reattached, the missing frames will produce abrupt jumps. Although a tear can physically be repaired, with the ends accurately reattached, the tear ends remain visible [Fos18]. Figure 3.9e shows an example of this artefact.

All of the above-described artefacts and defects can either be directly embedded in the



(a) Scratch (b) Cue dots (c) Low contrast

(d) Melt (e) Tear (f) Wrong exposure

Figure 3.9: Artefacts introduced from repeated replication and presentation.

filmstrips or the result of duplication and multiple generation copying [Fos18]. Since the replication process was not always executed with the appropriate equipment or in an ideal environment, artefacts like dirt and scratches often got copied directly into the films. Hence, with each replica, the films accumulated new defects and artefacts which introduced various distortions in the films. Furthermore, with every filmstrip copy, the quality of the material decreases. As a result, films of the two datasets are characterized with significantly low contrast and colour fades, as shown by the examples in Figure 3.9c and Figure 3.9f.

Faulty and inadequate film replication can be the root cause of more complex film damage and artefacts. Shaking or instability of the film is such an example. By copying incorrectly aligned filmstrips, faulty film replication causes frame displacements [ZMB12]. These frame displacements in combination with the displacements and distortions caused by the shrinkage of the filmstrips lead to the artefact nowadays known as instability [ZMB12]. Due to the utilization of manual cameras, historical films originally come with a certain instability level [Fos18]. Over the years, the instability of historical films is magnified through replication. The instability of historical films significantly inhibits any form of automatic video processing, especially motion tracking [ZZMB10].

Flicker is a similar defect case that often occurs as a result of copying errors. Essentially, flickering deals with unstable lighting of scenes [Fos18]. It is described as the effect that occurs as a result of the projection of a frame sequence with varying lightning [ZMB12] [Fos18]. Flicker is also caused by the film transports, which in the early filmmaking era was performed manually [ZMB12]. The manual film transport

(a) Underexposure (b) Overexposure (c) Flicker

Figure 3.10: The effects of wrong exposure and flicker.

resulted in a variable exposure time of the filmstrip. This, in turn, causes a difference in the brightness of the frames. In the literature, such variance in the brightness of the frames is referred to as the overexposure i.e. underexposure error (see Figure 3.10a and Figure 3.10b) [ZMZB11]. The historical films used in this work are heavily affected by flicker (see Figure 3.10c).

3.3.3 Challenges

All of these artefacts found in archive films are what distinguishes historical film material from contemporary or state-of-the-art film material. The above-described artefacts i.e. traits of historical films, however, significantly hinder the automatic video analysis process. Specifically for the problem of SBD, each type of artefact comes with a separate set of challenges. Moreover, the combination of multiple artefacts gives rise to a number of new problems and questions which current SBD methods simply cannot provide an answer to. The artefacts found in historical films used in this work undoubtedly interfere with the task of SBD and thus significantly increase the complexity of the SBD problem.

Artefacts like dirt, dust, scratches, mould and liquid spills present false and ambiguous information that negatively affect the SBD process [SZMB11] [ZMZB11] [ZMB12]. This kind of artefacts produces abrupt disturbances in the visual content of the frame. Since the general SBD pipeline is based on the extraction of visual features and their framewise comparison, the presence of such artefacts leads to a high number of false-positive cases of ATs. Furthermore, the effects of filmstrip shrinkage (i.e. the visible frame lines, borders and in some cases the partial visibility of the contents of the next shot) add unnecessary noise to the (dis-)similarity measure used for the comparison between frames [ZMZB11]. The scope of this subset of artefacts is a part of a frame, due to which they are referred to as local [ZMB12]. Consequently, local artefacts severely affect local features.

Artefacts that discontinue or interrupt the temporal axis of a film are known as temporal artefacts [ZMB12] [Fos18]. An example of a temporal artefact is a tear that results in an abrupt jump of the film. This kind of abrupt jumps interrupts the motion of a film. As a result, temporal artefacts represent a major problem for motion-based features and SBD methods as well as motion analysis and tracking [SZMB11].

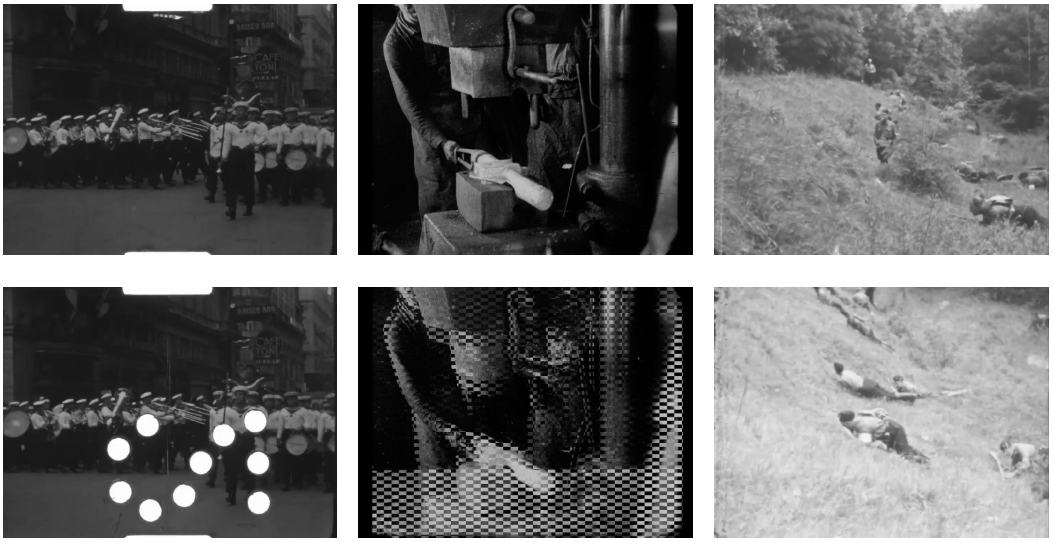
Figure 3.11: Blurred frames

In contrast to local artefacts, the effect of global artefacts targets the area of the whole frame [ZMB12]. Global artefacts include film instability, low-contrast and flicker [ZMB12]. Film instability significantly impeds the process of motion analysis and puts SBD methods that rely on motion features at a great disadvantage [ZZMB10]. Furthermore, the shaking results in noisy motion estimates which makes the tracking of objects in historical videos extremely difficult. On the other hand, flicker is critical as it alters and affects the brightness of the frames [Fos18]. This interferes with colour-based frame similarity methods which are vital for SBD and automatic video content-based retrieval. As a result, brightness information of historical film material is highly unreliable and nonrepresentative. This also true for the low-contrast artefact. Colour features ought not to be used in combination with low contrast frames as they will provide no explicit and representative information [ZMB12].

Apart from the abundance of various artefacts, the two historical film material datasets EFilms and IMC contain a large number of blurred frames. As shown by Figure 3.11, blurred frames are characterized by obscuring objects and edges and an overall inconcrete and non-distinct structure. This, causes problems as the visual features extracted from blurry frames have little expressive and representation power. In contrast, the opposite is true for sharp frames. As a result, the comparison of the visual features of a blurred frame with the visual features of a very similar sharp frame produces a high dissimilarity value. In terms of the task of SBD, blurry frames lead to a massive number of false-positive detections of ATs. This drastically decreases the performance rate of existing SBD approaches.

From a practical point of view, it is not always straightforward to determine whether to frames are a part of the same shot. The occurrences of many such cases are also evident in the films investigated in this work. This is due to the fact that the films share the same theme and stem from the same time period. Hence, the films often depict very semantically similar scenes. Furthermore, historical films from the EFilms and IMC datasets have a similar setting in terms of location and background. The challenges of identifying shot correspondence in historical films are captured by Figure 3.12. Figure 3.12c illustrates an example of similar two frames with the same context and location. The frames, however, belong to different shots. On the other hand, Figure 3.12a demonstrates two identical frames which belong to the same shot. These two frames only seem different due to the

artefacts found in the second frame. Additionally, the frames illustrated in Figure 3.12b also belong to the same shot. However, this is not immediately apparent due to the heavy visual degradations in the second frame. Ultimately, the perceptual similarities between the shots of historical videos contribute to and increase the complexity of determining the boundaries between the shots.



(a) Frames from identical shots including artefacts.

(b) Identical shot with complex film distortions.

(c) Frames with similar setting, but different shot.

Figure 3.12: Identical Shots vs Similar Shots.

These artefacts and challenges of historical films interfere with the performance of existing SBD methods. Nonetheless, established SBD methods are neither developed to support grayscale films nor optimized to mitigate the effects of the artefacts found in the historical film material [ZMB08] [ZMB12]. This makes existing SBD approaches inapplicable and unsuitable for videos and films from the historical domain. To counteract the challenges and artefacts of historical films, this thesis proposes and implements an SBD framework designed specifically to target historical films.

CHAPTER 4

Methodology

4.1 System overview

The proposed framework is based on deep learning and is inspired by state-of-the-art SBD methods. Figure 4.1 presents a visual representation of the proposed framework. The input of the proposed framework is a video file. The output of the framework, on the other hand, is a list of the start and end frame positions of the detected transitions. Fundamentally, the framework consists of three core components including an AT detector, and two specialized GT detectors: a dissolve detector and a FOI and wipe detector. Additionally, the framework incorporates a GT candidate selection module which is also an integral part of the pipeline.

The AT detector component is responsible for performing the detection of ATs in videos. The AT detection process follows the steps of the general SBD framework. As illustrated in Figure 4.2, the AT detection process involves the extraction and representation of frame content in the form of features, similarity calculation and transition classification.

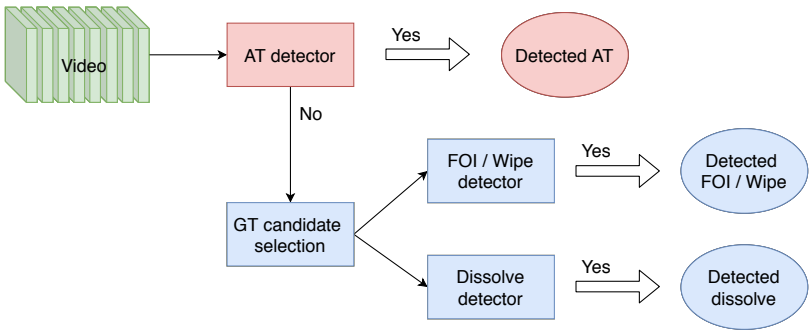


Figure 4.1: System overview of the proposed SBD framework.

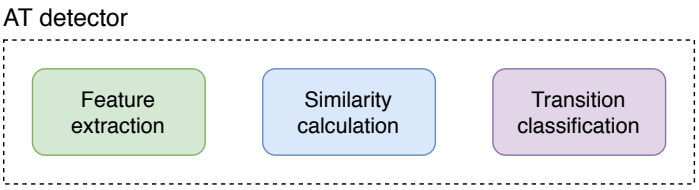


Figure 4.2: The main tasks of the AT Detector.

The AT detector employs a self-designed CNN model for the feature extraction and the similarity calculation. Essentially, the CNN model takes two frames (an image pair) as input and outputs their similarity value. This master thesis experiments and analyzes the effect of different backbone architectures which results in two CNN models called ATNet and ResidualATNet. Additionally, it investigates two feature extraction strategies: single feature extraction and multi-feature extraction. In the transition classification step, the AT detector compares the similarity value obtained from the CNN model against a threshold and declares either a transition or non-transition respectively. This thesis examines and tests a *fix*, as well as an *adaptive threshold* policy for the declaration of transitions. The frame pair which responds negatively to the AT detector (i.e. non-transition is declared) is passed to the *GT Candidate Selection module*. Further details about the AT detection process can be found in Section 4.3.

The GT Candidate Selection module serves the purpose of producing sets of specific types of candidate GT segments (see Figure 4.3). The frames which are forwarded to the GT Candidate Selection module go through a set of conditions and checks. The frames whose similarity value satisfies the dissolve condition form the set of Dissolve GT candidates. On the other hand, the frames whose similarity value satisfies the FOI / Wipe condition build the set of Wipe and Fade Out/In GT candidates. Otherwise, the frames are discarded if their similarity value fulfils none of the conditions. Neighbouring frames in the formed GT candidate sets (Dissolve and FOI / Wipe) are concatenated together to form GT candidate video segments. The Dissolve GT candidate video segments and the FOI / Wipe GT candidate video segments are then passed to the corresponding detector for GT detection.

The FOI / Wipe detector and the Dissolve detector differ in the type of GT candidate video segments they receive as input. Nonetheless, the background processes performed by the two detectors are similar and are summarized in Figure 4.4. In order to be able to perform GT detection, the two detectors rely on the same CNN model called ResidualGTNet. This ResidualGTNet model takes a sequence of frames as input and outputs a similarity matrix. After extraction of the features of the input frames the model creates a similarity matrix. The matrix contains the similarity values of all possible feature comparison combinations between the sequence of frames under consideration. From the similarity matrix, the average value of the similarities above the main diagonal is computed. To declare and adjust the start and end position of a GT, the FOI /

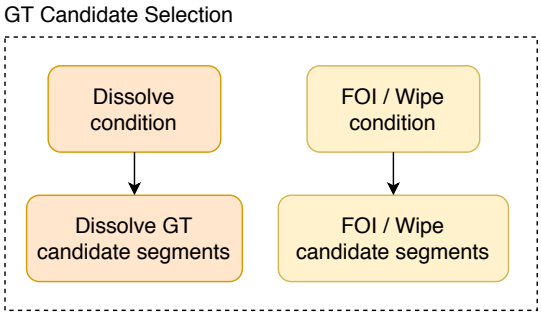


Figure 4.3: The GT Candidate Selection module produces two sets of GT candidate segments.

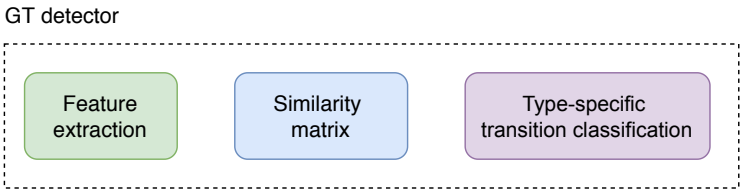


Figure 4.4: The main processes performed by the GT detectors.

Wipe detector and the Dissolve detector then compare this value against thresholds and conditions specific for their respective GT type. The details about the GT detection algorithm are provided in Section 4.4.

4.2 Data Engineering

The current state-of-the-art SBD methods which rely on deep learning models are trained and evaluated on contemporary video datasets such as RAI [BGC15b], BBC Planet Earth [BGC15a] and ClipShots [TFK+18]. However, to provide support for grayscale historical films, training and learning from historical data is of crucial importance. To improve the overall SBD performance in the historical domain, it is required that the CNN models utilized in the proposed SBD framework are trained on historical video data. Due to the insufficient digitization, controlled access and sensitivity of the content found in historical films, there is an essential lack of historical video datasets on the Internet [ZMB08]. For this reason, a new self-designed dataset, called *HistoricalDataset* is created.

For the creation of the HistoricalDataset sample frames from both historical film datasets, EFilms and IMC are used. A total of 12 historical films (6 from EFilms and 6 from IMC) are used to create the HistoricalDataset. These films are excluded from the test dataset

films which are used for the evaluation of the CNN models. Furthermore, out of the 12 selected films, 8 films (4 from EFilms and 4 from IMC) are used to create the training samples and the rest are utilised for the creation of the validation samples.

A sample of the HistoricalDataset is represented by an image pair and an integer label. The label indicates whether a transition between the two images exists. The value of the label can be either 0 meaning no transition exists (i.e. the image pair belongs to the same shot) or 1 meaning an AT exists between the image pair. The pairs which represent a transition are artificially created, meaning the transition does not exist in any video from the EFilms and IMC datasets. Figure 4.5a and 4.5b show examples of samples that have label 0 and hence the images belong to the same shot. In contrast, positive samples that contain ATs are illustrated in Figure 4.5c and 4.5d.

(a) Negative sample (b) Negative sample (c) Positive sample (d) Positive sample

Figure 4.5: Samples of the *HistoricalDataset*.

The HistoricalDataset contains a total of 3298 samples. The dataset is carefully created with an approximately even class distribution between positive and negative samples. Out of the 3298 samples, 1648 samples are negative (do not contain a transition and have a label 0) whereas 1650 samples are positive (contain an AT and have label 1). Furthermore, the diversity of the data is ensured by utilizing frames from both datasets EFilms and IMC. Fifty per cent of the negative samples are taken from shots which belong the EFilms dataset, whereas the other half of negative samples stem from shots from the IMC dataset. To avoid overfitting, half of the positive samples contain one frame from the EFilms dataset and one frame from the IMC dataset. Consequently, the two datasets EFilms and IMC are not only equally represented in the HistoricalDataset but they are also equally distributed through the positive and negative samples of the new dataset.

The samples used for the training of the CNN models correspond to 70% of the HistoricalDataset. The training dataset contains 2299 samples, 1149 of which do not contain a transition (i.e. label 0) and the rest of the 1153 samples contain an AT (i.e. label 1).

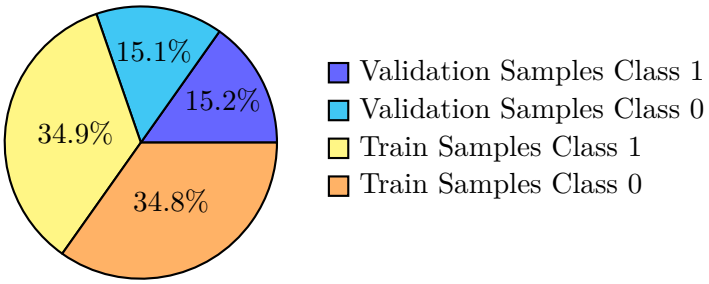


Figure 4.6: Separation of the *HistoricalDataset* into Train and Validation samples in percentage.

	No. of Films	Transitions	Non - transitions	Total
Train	8	1153	1149	2299
Validation	4	500	499	999

Table 4.1: Separation of the *HistoricalDataset* into Train and Validation sets.

Out of the 2299 training samples, 965 samples stem from the 4 EFilms films, 830 samples are created from the other 4 IMC films and 504 samples (label 1) contain one frame from the EFilms films and the other from the IMC films.

The other 30% of HistoricalDataset are used for validation of the CNN models. The validation dataset contains 999 samples, 499 of which do not contain a transition (i.e. label 0) and the other 500 contain an AT (i.e. label 1). Similarly to the training dataset, the 999 validation samples are split equally between the films of EFilms (428 samples from 2 historical films) and IMC (401 samples from 2 historical films) and contain 170 samples which have one frame from the EFilms and the other from the IMC films. The pie chart in Figure 4.6 illustrates the separation of the HistoricalDataset in percentage. Furthermore, Table 4.1 shows the split of the dataset in terms of the total number of samples and the number of positive and negative samples contained in each subset.

The performance of the CNN models is evaluated on separate 115 test films, 55 of which are from the EFilms dataset and the other 60 from the IMC dataset. Table 4.2 shows the statistics of the test films. The test historical films from EFilms contain a total of 6501 transitions, whereas the test films from IMC contain 4514 transitions.

In addition to the HistoricalDataset which has three colour channels (RGB), a Hisorical-DatasetGrey is generated. The HisoricalDatasetGrey contains the same dataset samples as the original, however, the images are converted to a grayscale format. The new dataset is created because most of the historical films are grayscale.

Dataset	No. of Films	Transitions
Test_EFilms	55	6501
Test_IMC	60	4514
Total	115	11015

Table 4.2: Test datasets used for the evaluation of the CNN models.

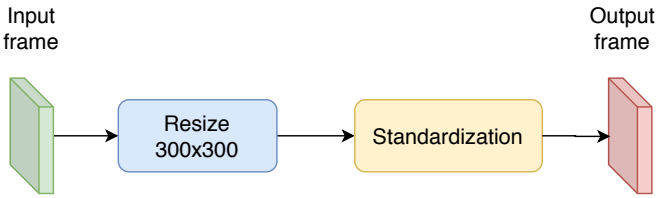


Figure 4.7: Data Preprocessing Pipeline.

4.2.1 Data Preprocessing

The first step in training, before feeding the input data in the deep neural network is preparing and preprocessing the data [PS16]. While image pixels are often integers in the range from 0 to 255 and can be fed raw to a CNN model, preprocessing the data has advantages such as data consistency and a faster and more reliable training [PS16].

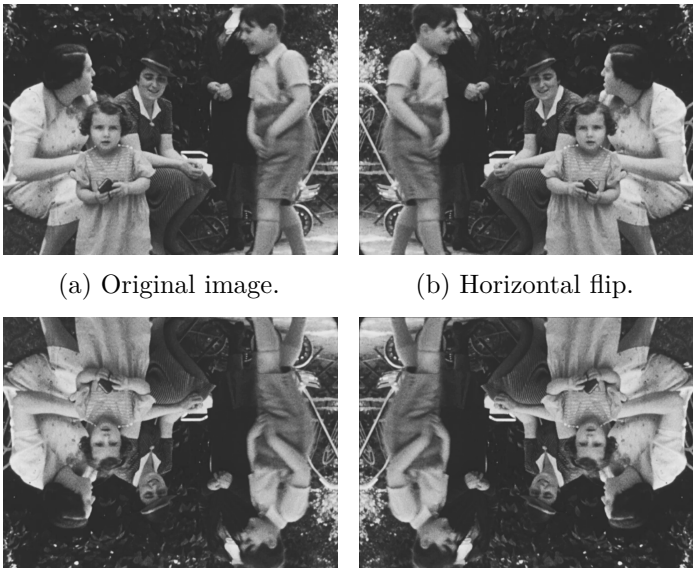
Before the start of the training process, each frame passes through a predefined preprocessing pipeline. A visual illustration of the preprocessing pipeline is depicted in Figure 4.7. In the first step of the preprocessing pipeline, the frame is resized to 300x300 pixels. The resized image is then forwarded to the next step - standardization. To ensure the consistency of the input data, the frame is then standardized globally across every channel. Hence, the process of standardization includes the precalculation of the mean and standard deviation values of each (RGB) channel of the frames from the training dataset [PS16]. The standardization of the frame is performed according to Equation 4.1 [PS16]:

$$img = \frac{(img - mean)}{std} \tag{4.1}$$

where img represents the frame and the mean and std represent the precomputed values of the training dataset. This step concludes the preprocessing pipeline and the resulting frame is ready to be used in by the CNN model.

4.2.2 Data augmentation

The efficiency and performance of CNN models depend on the quality of the data [PS16]. Consequently, it is of vital importance that the data utilized is diverse, and satisfies



(a) Original image. (b) Horizontal flip.

(c) Vertical flip. (d) Horizontal + Vertical flip.

Figure 4.8: Data augmentation techniques applied to the *HistoricalDataset*.

both the qualitative and quantitative requirements of the specific task. The method of data augmentation is useful in cases where the data sources are limited or the process of data gathering proves to be difficult and time-consuming. Data augmentation artificially increases an image dataset of limited size by applying various transformations to the existing images of the dataset [KSH12]. Commonly, the most basic transformations performed to augment an image dataset include cropping, scaling, flipping, rotation or any combination of them [KSH12]. Moreover, data augmentation helps CNN models to become more robust and ultimately increases their invariance to rotation, size and illumination. The data augmentation techniques can either be performed during the data preprocessing step or they can be applied during the training process on the fly [KSH12].

Data augmentation techniques are applied to the HistoricalDataset in order to improve the rotation and flip invariance of the CNN models utilized in the proposed SBD framework. The HistoricalDataset is augmented using a set of three types of transformations: horizontal flip, vertical flip and a combination (horizontal + vertical flip) of the two. Examples of the pre- and post-application effects of these transformations are depicted in Figure 4.8. The set of these data augmentation techniques are performed on the fly and increase the original HistoricalDataset by 20%. The augmentation techniques are performed on randomly selected samples from the dataset. Since a sample in the case of the HistoricalDataset includes two images (see Figure 4.5), the transformation is applied to both images.

4.3 Abrupt Transition Detection

4.3.1 CNN Architectures

The main tasks and responsibilities of the CNN models consist of extraction of high-level CNN features and calculation of similarity between the extracted features. The proposed models are fully convolutional which means they include convolutional layers and no fully connected layers. The output of the networks represents a similarity value for a given input image pair.

The basic architecture of the CNN models is inspired by and resembles a Siamese network [GFZ+18]. Hence, the networks take as input a pair of images of size 300 x 300 pixels. The images are processed in parallel by the separate feature extractor branches. The feature extractor branches have the same architecture and share the same weights. Moreover, for each image, the feature extractors extract information from the convolutional layers in the form of feature vectors. The extracted features are used for the similarity calculation between the two input images. The final output of the CNN model represents a similarity score which is computed using cosine similarity.

This thesis investigates the effects of different feature extractors on the efficiency and overall AT detection performance. For this reason, two different feature extractor models are utilized and examined. This results in two separate CNN architectures for AT detection called: ATNet and ResidualATNet.

ATNet

The complete architecture of ATNet is demonstrated in Figure 4.9. The input of ATNet is an image pair which is simultaneously processed by the feature extractor branches. ATNet has 26 convolutional layers and 10 pooling layers in total. Additionally, ATNet utilizes L2Normalisation layers and Flatten layers. The L2Normalisation layers are used to normalise the values of the input vector in the range between 0 and 1.

The feature extractor of ATNet is inspired by the state-of-the-art neural network VGG16 [SZ15]. It contains 13 convolutional layers and 5 pooling layers (see Figure 4.9). Essentially, the architecture of the feature extractor is separated into five blocks of convolutional layers with a kernel size of 3x3. Each convolutional layer is followed by an activation layer of type ReLU. Furthermore, each convolutional block is followed by a pooling layer of type max pool. The first two blocks contain two convolutional layers. The blocks are followed by a pooling layer with a kernel size of 2x2. The depth hyperparameter (i.e. the number of convolutional kernels per layer) of the convolutional layers in the first two blocks is set to 64 and 128 respectively. The structure of the other three blocks consists of three convolutional layers. Similarly to the first two blocks, each convolutional block is followed by a pooling layer of type max pool and a kernel size of 2x2. The number of convolutional kernels used in the last three blocks of convolutional layers is 256, 512 and 512 respectively. At the very end of the feature extractor, the

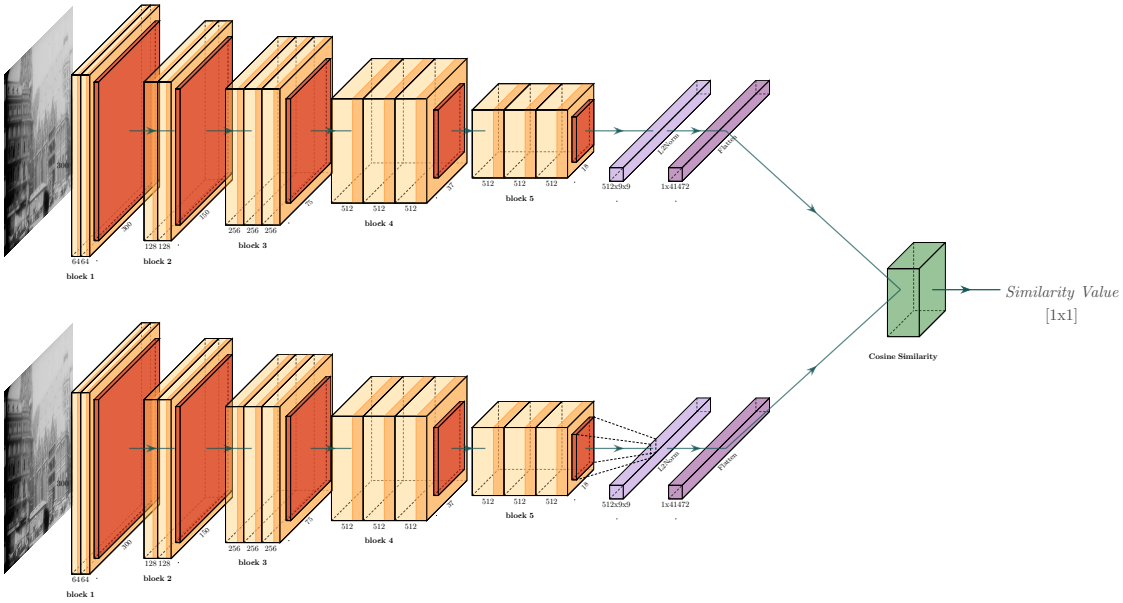


Figure 4.9: Detailed architecture of ATNet.

output is normalized and flattened. The final output of the model is a number between 0 and 1 which denotes the similarity between the two input images.

ResidualATNet

The feature extractor of the ResidualATNet is inspired by the state-of-the-art CNN architecture of ResNet18 [HZRS16]. Therefore, the layers of the ResidualATNet model rely on learning residual functions with a reference to the input layer. In other words, if the underlying function is $H(x)$, then a stack of non-linear layers are allowed to fit another function mapping of the form: $F(x) = H(x) - x$. This allows the original underlying function $H(x)$ to be recast in a residual form, namely $F(x) + x$. The residual $F(x) + x$ operation is executed by a so-called shortcut connection and an addition operation [HZRS16]. Hence the name of the CNN model ResidualATNet.

The feature extractor starts with a convolutional layer with a kernel size of 7x7. This layer is followed by a max-pooling layer with a kernel size of 3x3 and a stride of 2. The feature extractor consists of an additional 16 convolutional layers with kernels of size 3x3. Furthermore, it utilizes ReLU as an activation layer and employs Batch normalization layers to normalize the output of the convolutional layers. The convolutional layers are structured into 4 blocks each containing 4 convolutional layers (see Figure 4.10). The core components in each block are the residual shortcut connections. This means that each block is built from a projection shortcut connection and an identity shortcut connection [HZRS16].

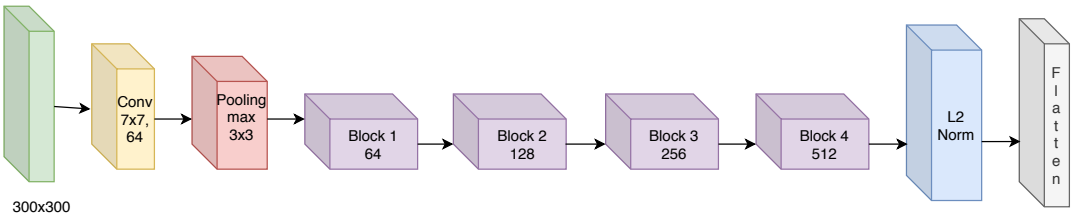
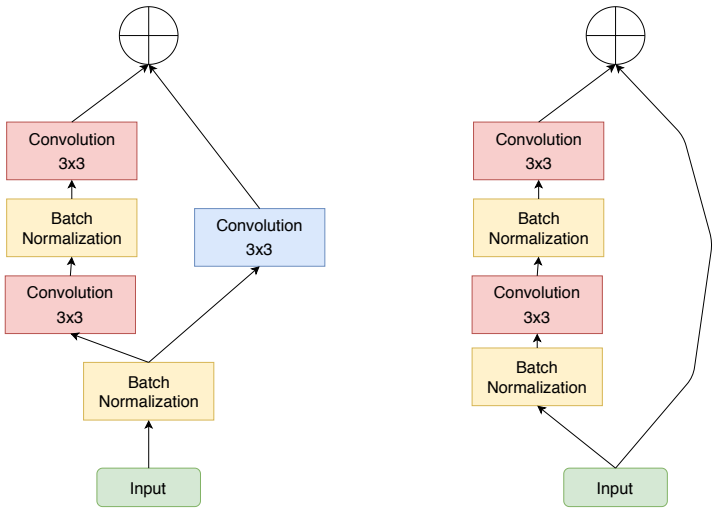


Figure 4.10: Feature Extractor of ResidualATNet.



(a) Projection shortcut connection. (b) Identity shortcut connection.

Figure 4.11: Shortcut connections of ResidualATNet. [HZRS16]

Figure 4.11 illustrates the structure of a projection and an identity shortcut connection. The identity shortcut connection is also called a skip connection due to the fact that it skips through one or more stacks of convolutional layers [HZRS16]. The identity shortcut connection simply passes the input volume to the addition operator. Such connections can be used directly if the input and output are of the same dimensions. However, if a dimension mismatch occurs between the input and the output volume, projection shortcut connections must be utilized. In a projection shortcut connection, there is an additional convolutional layer with a kernel size of 1x1 and a stride of 2 [HZRS16]. The convolutions in projection shortcut connections are used to match dimensions i.e. downsample the volume before the execution of the addition operation [HZRS16]. The residual connections improve the performance and solve the problem of the vanishing gradient by allowing a fast backward propagation in the network through the skip connections [HZRS16].

Similarly like ATNet, the input of the full ResidualATNet model is an image pair of

size 300x300 pixels. The images are then forwarded to the separate feature extractor branches for further processing. After the features of the images are extracted, both of the features are normalized and flattened. The normalization transforms the range of the feature vectors between 0 and 1. Lastly, the similarity between the features is computed which represents the output of the model.

Similarity Calculation

The similarity calculator module is used by the ATNet and ResudualATNet and is responsible for computing the similarity between the two input images. In both of the models ATNet and ResidualATNet, the calculation of the similarity is performed in an identical way using Cosine similarity.

The Cosine similarity represents a popular measure of similarity between two non-zero vectors [XSX16] [TFK+18]. It is defined as the cosine angle between the two feature vectors in a multi-dimensional space [XSX16]. In other words, this is the same as computing the inner product between two normalized vectors. The Cosine similarity is advantageous over Euclidean distance because even if two feature vectors are apart in terms of Euclidean distance, the angle between them could still be small [XSX16]. In the context of Cosine similarity, the smaller the angle, the higher the similarity between the two vectors.

Mathematically, the Cosine similarity of two feature vectors f_i and f_j is defined as [XSX16]:

$$cos(f_i, f_j) = \frac{f_i f_j}{\|f_i\|\|f_j\|} = \frac{\sum_{k=1}^{n} f_{ik} f_{jk}}{\sqrt{\sum_{k=1}^{n} (f_{ik})^2}\sqrt{\sum_{k=1}^{n} (f_{jk})^2}} \tag{4.2}$$

Where f_{ik} and f_{jk} are vector components respectively. Prior to computing the Cosine similarity, the feature vectors are normalized so their values range between 0 and 1. As a result, the angle between such feature vectors ranges between 0 and 90 degrees. While an angle of 0 degrees indicates the greatest similarity between the feature vectors, an angle of 90 degrees represents complete decorrelation. Consequently, the Cosine similarity value for such vectors ranges between 0 and 1.

The term Cosine distance represents the complete opposite of Cosine similarity. The formulation of Cosine distance presented in Equation 4.3. The idea behind this is that if two vectors are identical then their Cosine similarity is 1. Therefore, their Cosine distance equals 0.

$$CosineDistance = 1 - CosineSimilarity \tag{4.3}$$

In summary, the similarity calculation component in the two models ATNet and ResidualATNet initially performs a dot product operation between the two normalized feature vectors. This results in the Cosine similarity. For the training process, the similarity value is transformed into a Cosine distance using the subtraction operation.

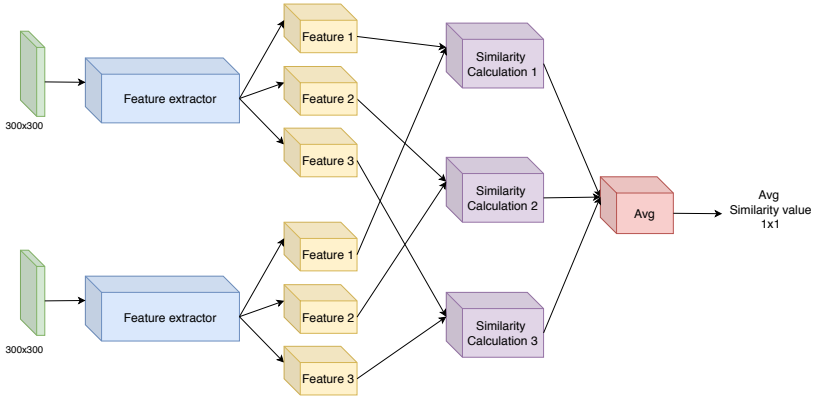


Figure 4.12: CNN Architecture using 3 feature vectors for the similarity calculation.

Multi-feature information

In CNN models, each CNN layer outputs a feature map which is the result of applying filters to the input image [LBD+89]. The feature map captures significant information specific to the input image. Commonly, the early layers of a CNN model that are closer to the input maintain the most information about the input [KSH12]. Furthermore, they detect and capture low-level features such as colour and edges. Going deeper into the CNN model the features become more abstract. The later layers of the CNN model capture high-level features used to detect objects and shapes [LBD+89]. Therefore, the deeper features of the network are the most complex and provide information relevant to the output of the network [LBD+89] [KSH12].

In the original configuration of the ATNet and ResidualATNet models, the last feature layer is used for the similarity calculation between the input image pair (see Figure 4.9). Furthermore, this master thesis experiments and examines the effects of multi-feature information on the similarity calculation. The idea behind this lies in the fact that heterogeneous feature information (of both higher and lower level) can provide a more genuine similarity value for the input image pair. To leverage the extraction of a combination of features, two new CNN inference architectures for ATNet and ResidualATNet are created. These architectures extract feature information from 3 layers as shown by Figure 4.12. In the ATNet model, the feature information from the 7th, 10th and 13th convolutional layer is extracted. Moreover, the ResidualATNet model extracts the features from the convolutional layers 7, 11 and 15. The extracted feature vectors are forwarded to the similarity calculation module. The output of this architecture (see Figure 4.12) represents the average similarity value between the extracted features.

4.3.2 Training

The CNN models ATNet and ResidualATNet are implemented in the MXNet [CLL+15] framework. An MXNet model contains two files: a JSON file and a parameter file [CLL+15].

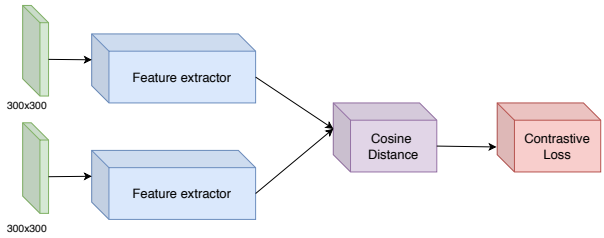


Figure 4.13: Training architecture of ATNet and ResidualATNet.

The JSON file represents the structure of the model and the parameter file specifies the weights of the model. ATNet and ResidualATNet are trained on the HistoricalDataset and the HistoricalDatasetGrey. Furthermore, each of the models is trained with the application of data augmentation techniques on the fly. An abstract training architecture of the proposed CNN models is presented in Figure 4.13.

Weight initialization. First, the weights of the models are initialized. Since both ATNet and ResidualATNet are inspired by state-of-the-art architectures transfer learning is utilized [SZL15]. The weights of ATNet are initialized with the weights of the model VGG16 [SZ15] pre-trained on PASCAL VOC 2007 [EGW+10] and PASCAL VOC 2012 [EGW+10]. Similarly, the weights of ResidualATNet are initialized with the weights of the ResNet [HZRS16] model pre-trained on PASCAL VOC 2007 [EGW+10] and PASCAL VOC 2012 [EGW+10].

Hyperparameters configuration. The initialization of the hyperparameters is critical for obtaining a high-performance model [Ben12]. For the training of ATNet and ResidualATNet, the Stochastic Gradient Descent (SGD) [BL03] is employed as the main optimization algorithm. The SGD algorithm updates the parameters and learns on each training sample [BL03]. Furthermore, the batch size utilized for the training process is configured to 64.

Apart from the optimization algorithm, another key hyperparameter that influences the behaviour of learning is the learning rate [Ben12]. The learning rate controls the speed of learning and consequently affects the velocity of convergence [BL03] [Ben12]. The learning rate can either have a fixed value throughout the whole training process or change according to a specific strategy [Ben12]. Schedulers leverage the possibility of altering the learning rate and offer several strategies. A MultiFactor Scheduler is employed for the training of ATNet. The MultiFactor Scheduler follows a stepwise decay strategy in which the learning rate is decreased by a factor of 10 at a specific interval. The models are trained with a learning rate of 0.001 for the first 100 epochs. The learning rate is then decreased by a factor of 10 for the remaining 50 epochs.

Before the training process begins, all images are standardized using the mean and standard deviation values of the RGB channels (meanR, meanG, meanB, stdR, stdG and stdB) of the training dataset. These values are pre-computed for the two training

Dataset	meanR	meanG	meanB	stdR	stdG	stdB
HistoricalDataset	82.5	84.16	83.22	63.76	63.39	63.29
HistoricalDatasetGrey	83.56	83.56	83.56	63.32	63.32	63.32

Table 4.3: Global mean and standard deviation values of the image channels across the train datasets.

datasets HistoricalDataset and HistoricalDatasetGrey separately. The corresponding mean and standard deviation values are provided in Table 4.3. Finally, the weight decay parameter is initialized with a value of 0.0005.

Loss function. Contrastive loss is utilized as a loss function for the training of the CNN models. The contrastive loss is selected for the training as it operates on a sample pair [HCL06]. Developed by Hadsell et al. [HCL06], this loss function is particularly used for the problem of image similarity calculation. Intuitively, the main idea behind the contrastive loss function is to place similar pairs closer together and to push apart dissimilar pairs [HCL06]. In other words, the contrastive loss function tries to minimize the distance between similar pairs and maximize the distance between dissimilar pairs. Therefore, the loss value is lower for a similar pair and higher for a dissimilar pair. The exact definition of contrastive loss can be observed in Equation 4.4 [HCL06] where f_i, f_j are the feature vectors of the sample pair. The $l(i, j)$ is defined as a binary indicator of whether the pair (i, j) is to be considered similar. In our case, a label of value 1 indicates that an AT exists between the images (i.e. the image pair is dissimilar). Conversely, a label of value 0 represents a non-transition between the images (i.e. the pair is similar). Moreover, $D(f_i, f_j)$ represents the Cosine distance between the feature vectors f_i, f_j of the sample pair. Finally, m is defined as a positive margin which ensures that only samples of dissimilar pairs contribute to the loss if the distance is within the margin [HCL06].

$$loss(f_i, f_j) = \frac{1}{2}\left[(1 - l(i,j)) * D(f_i, f_j)^2 + l(i,j) * max(0, m - D(f_i, f_j))^2\right] \tag{4.4}$$

The graphs which contain the training and validation loss curves for each training process are depicted in Figure 4.14. The graphs demonstrate the values of training and validation loss over the epochs in the training process. As shown in the graphs the training process runs considerably smooth. In the first couple of epochs, a sharp decrease in both training and validation curves can be observed. In the rest of the epochs, the two losses continue to gradually decrease to 0.

When trained on the HistoricalDataset, the ATNet and ResidualATNet model reach a training loss of 0.001 and 0.015 respectively. The validation loss on the HistoricalDataset equals 0.007 for ATNet and 0.02 for ResidualATNet. In comparison to the Historical-Dataset, on the HistoricalDatasetGrey, ATNet and ResidualATNet produce a training loss of 0.002 and 0.011 and a validation loss of 0.008 and 0.016 respectively. Lastly, on

the HistoricalDataset + DA, ATNet achieves a training and validation loss of 0.002 and 0.008. The training and validation loss of ResidualATNet on the HistoricalDataset + DA have the same value of 0.011 and 0.016 respectively.

From the curves shown in Figure 4.14, it can be inferred that the training loss values are always lower than the validation loss values. Furthermore, the smoothness of the training curves indicates that the hyperparameters learning rate, weight decay and multifactor scheduler are properly configured. The validation curve and validation loss achieved by all models provide first results and indicate the behaviour i.e. performance of the models on unseen data. The tight gap between the train and validation loss curves shows that the learning process of the models runs as expected. This means the models are able properly to distinguish between transition and non-transition pairs.

4.3.3 AT Classification

In the next step, the predicted similarity values are compared to a specified threshold. If a similarity value is lower than the threshold an AT at the corresponding frame position is reported. Otherwise, the respective frame pairs and their similarity values are forwarded to the GT Candidate selection module for additional processing. Figure 4.15 depicts the flowchart of the complete AT detection process.

There exist different possibilities for dealing with a threshold. In this work, two types of threshold policies are implemented. The first one utilizes a *fixed threshold*. This means the value of the threshold does not change throughout the processing of a video dataset but remains fixed each video regardless of the similarity values. The second policy employs an *adaptive threshold*. The value of adaptive threshold changes and adjusts on the video processed. The utilization of an adaptive threshold is preferred, as this makes the complete detection method generalizable and independent of a hard-coded value.

The formula for calculation of the adaptive threshold, in this work, is inspired by the statistical outlier test called Chauvenet's criterion [Cha91]. In this case, for a list of similarity values, the application of Chauvenet's criterion returns a list of outliers [Cha91]. Figure 4.16 shows a visual illustration of the application of Chauvenet's criterion on two separate videos. Intuitively, the outliers in the case of AT detection represent sudden declines in similarity. Furthermore, the outliers admit lower similarity values and refer to transitions in between the frames. From the list of outlier similarity values, the average value is calculated. The average value is used as a threshold for the specific video. Finally, if a similarity value between two frames is below the computed threshold then an AT is declared.

4.4 Gradual Transition Detection

The GT Detection component of the proposed SBD framework explicitly targets and deals with the detection of GTs. The component has the ability to successfully detect and localize the three main GT types: dissolve, wipe and fade out/in transitions. The SBD

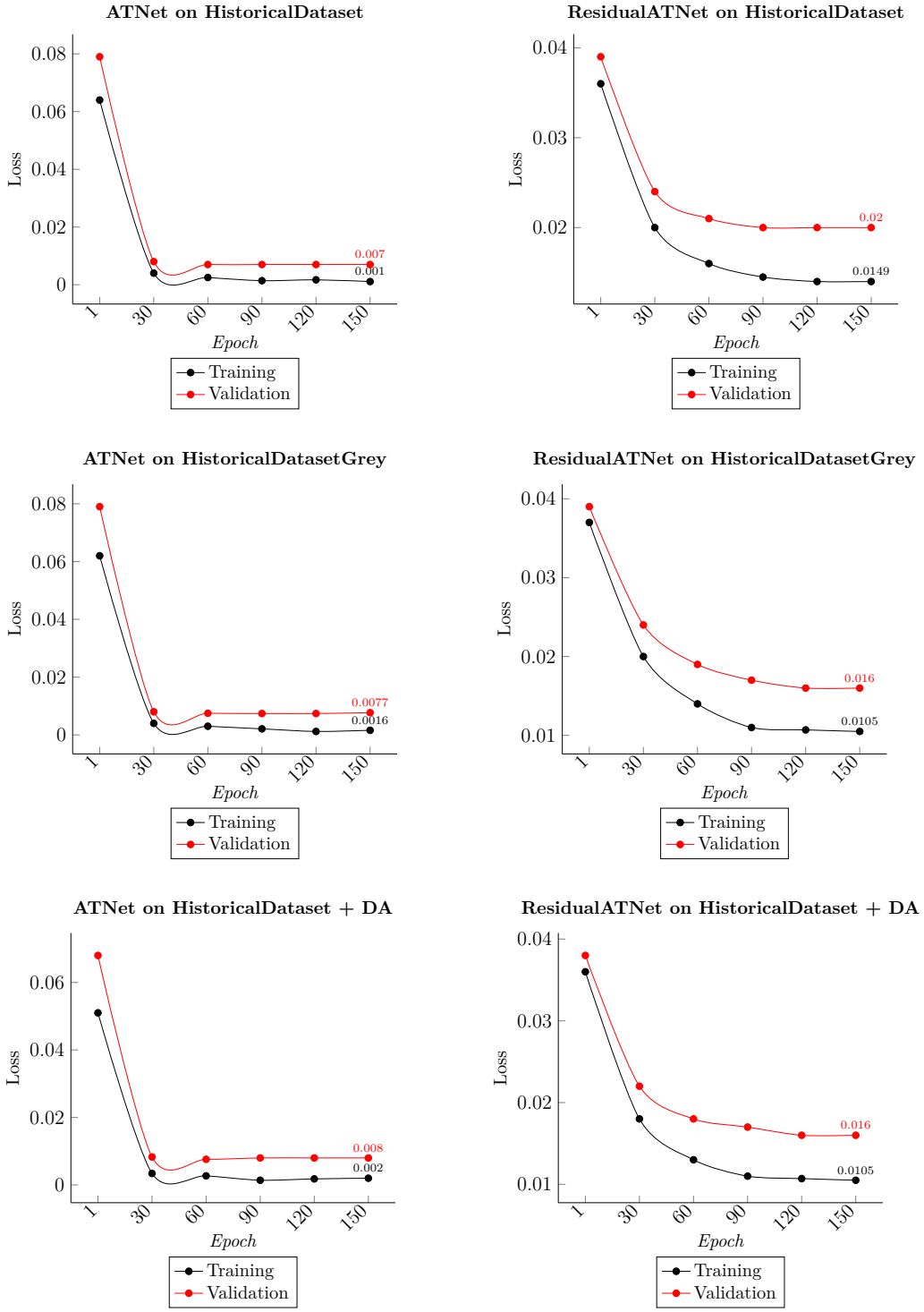


Figure 4.14: Train and Validation loss curves of the training processes of the different models.

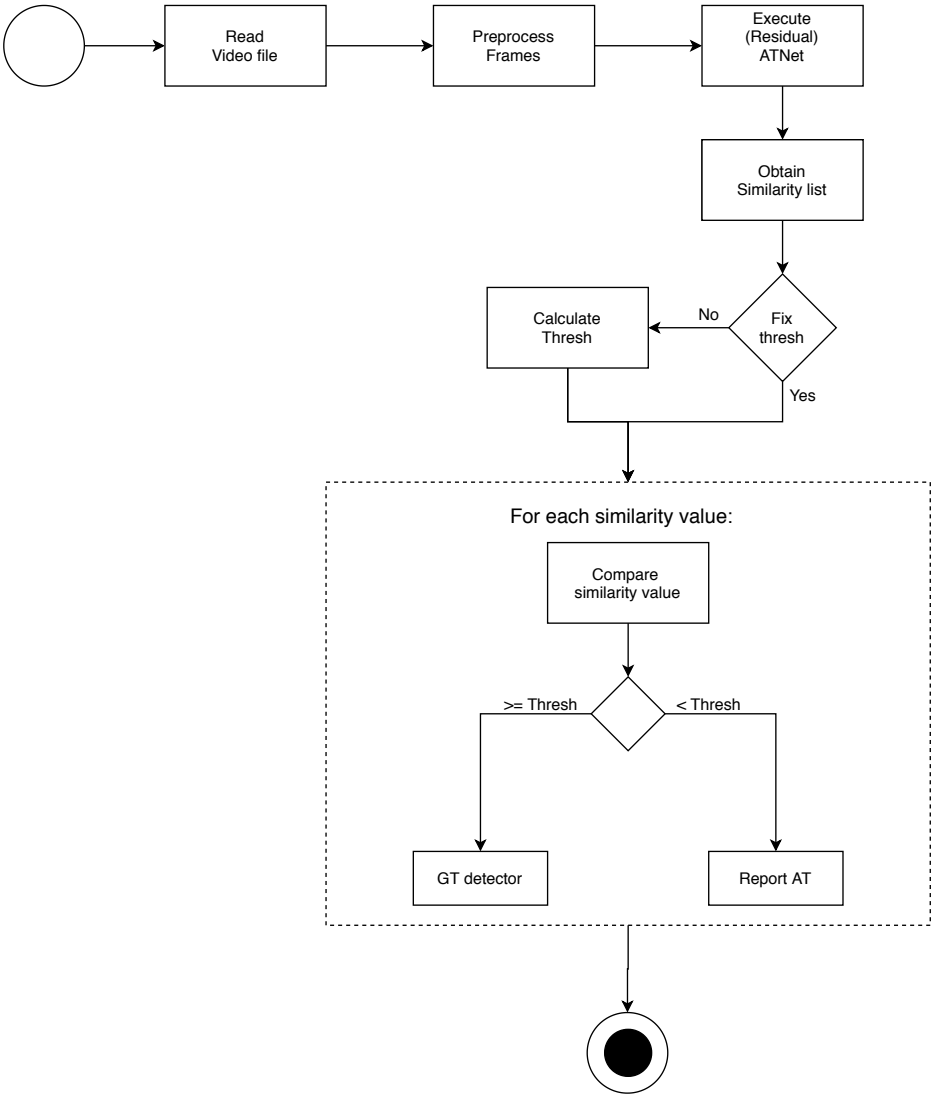


Figure 4.15: Flowchart of AT detection process.

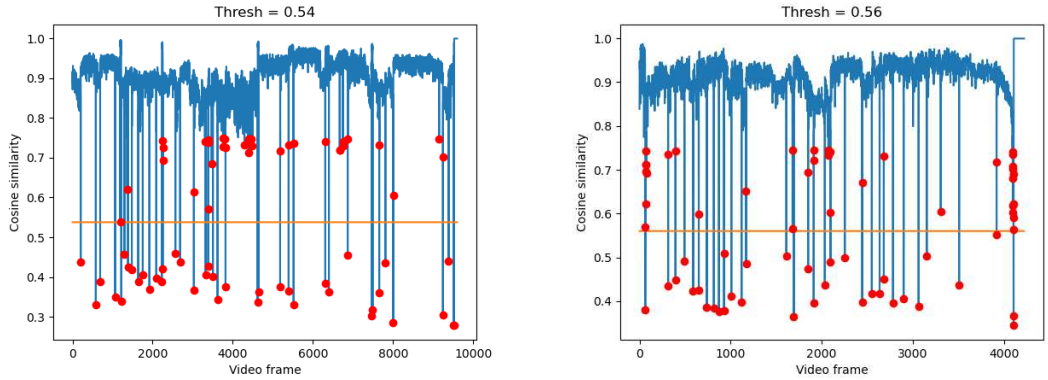


Figure 4.16: The model predicts a similarity value for each pair of adjacent frames in a video (blue line). The application of Chauvenet's criterion returns a list of outliers (red dots). The threshold is the mean of the outlier values (yellow line).

framework also includes a GT candidate selection module which produces a set of GT candidates. Two separate detectors perform the detection of GTs. While one detector concentrates solely on the detection dissolve transitions (Dissolve detector), the focus of the other detector lies in detecting wipe and FOI transitions (FOI / Wipe detector). Even though the two detectors endeavour to detect different GT types, they both utilize the same deep CNN architecture model for the detection. This section first explains the process of GT candidate selection and presents the architectural details of the CNN model used for GT detection.

4.4.1 GT Candidate Selection

The GT Candidate Selection module is responsible for producing sets of GT candidates. For this reason, the GT Candidate selection module introduces two selection conditions: a dissolve selection condition and an FOI/ wipe selection condition. The two selection conditions are defined in Equation 4.5 and Equation 4.6 where $sim(i, j)$ in these two equations represents the similarity value between two successive frames i and j. The input of the GT Candidate Selection module is represented by frame pairs which responded negatively to the AT detector and their similarity value. The frame pairs with similarity value that satisfies the dissolve selection condition are added to the set of dissolve GT candidates. Conversely, the set of FOI/Wipe GT candidates is formed by frame pairs that satisfy the FOI/ wipe selection condition. The frame pairs that neither satisfy the dissolve selection condition nor the FOI/ wipe selection are discarded.

$$\text{Dissolve selection condition: } sim(i, j) > S_d \tag{4.5}$$

$$\text{Fade / Wipe selection condition: } S_{min\_fw} < sim(i, j) < S_{max\_fw} \tag{4.6}$$

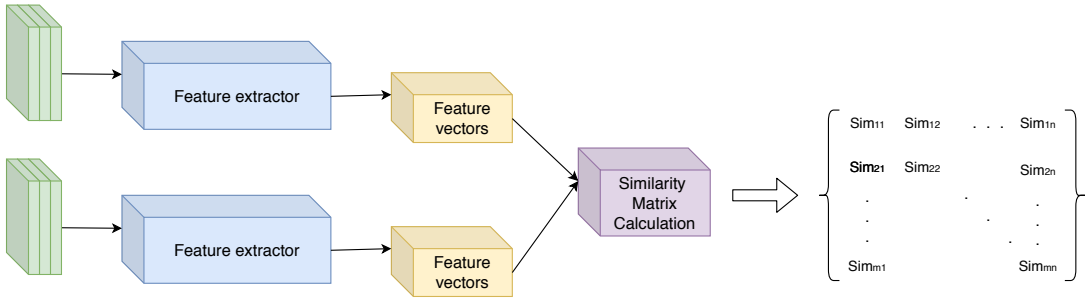


Figure 4.17: Architecture of ResidualGTNet.

Each set of GT Candidates is then prepared separately. The neighbouring frames in the GT candidate sets are concatenated together to form video candidate segments. Afterwards, the video segments which contain less than five frames are filtered out and removed from the GT candidate sets. The rest of the video segments are expanded by 20 frames. As such the Dissolve video segments and FOI/ Wipe video segments are individually forwarded to the CNN model called ResidualGTNet for further processing. After a number of empirical examinations, the values of parameters S_d, S_{min_fw} and S_{max_fw} in the GT selection conditions are set to 0.97, 0.7 and 0.93 respectively.

4.4.2 ResidualGTNet

ResidualGTNet represents a state-of-the-art deep CNN model which assists the detection of GTs by performing feature extraction and similarity computation. The input of ResidualGTNet is a sequence of frames. The length of the input frame sequence can vary and depends on the processing power of the GPU used. Overall, the structure and architecture of ResidualGTNet are very similar to the structure and architecture of ResidualATNet. ResidualGTNet utilizes the same feature extractor network as ResidualATNet which is presented and described in Section 4.3.1. Furthermore, for the extraction of features, ResidualGTNet utilizes the multi-information feature strategy and extracts feature vectors from three different layers. The specifics about the multi-information feature extraction strategy are provided in Section 4.3.1. The main difference between ResidualGTNet and ResidualATNet is the type of output the networks generate. Moreover, when executed on a video file, the output of ResidualATNet represents a list of similarity values between all consecutive video frames. In contrast, the output of ResidualGTNet is not a single value, but a matrix of similarity values. The structure of the similarity matrix together with the abstract architecture of the ResidualGTNet model can be observed in Figure 4.17.

The ResidualGTNet model creates the output matrix by computing the similarity values between all combinations of the input frame sequence (see Figure 4.18). The similarity matrix is calculated by using the dot product i.e. Cosine similarity between the feature

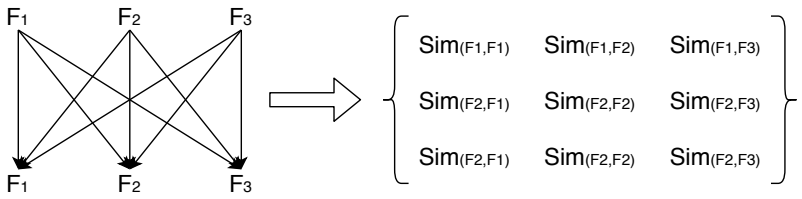


Figure 4.18: For a given input frame sequence $\{F_1, F_2, F_3\}$, the ResidualGTNet model outputs a matrix by calculating the similarity value between all input frame combinations.

vectors of the input frames. For example, the value at the matrix position $(1, n)$ represents the similarity value between the frame at position 1 and n respectively. Furthermore, the values of the main diagonal represent self-similarities and are thus characterized by the highest similarity values. Since the matrix is symmetrical, the similarity values above and below the main diagonal are identical. As a result, only the values above the main diagonal are considered relevant for the detection of GTs.

4.4.3 GT Classification

Each GT candidate segment produced during the GT candidate selection stage (see Section 4.4.1) is forwarded to the ResidualGTNet model. The ResidualGTNet model processes each GT candidate segment individually, with a batch size of 64 and outputs a similarity matrix. From the obtained similarity matrix, the average similarity value of the values above the main diagonal is calculated. To declare a GT, the average similarity value is tested and compared against different conditions and thresholds depending on the type of the GT candidate segment under consideration.

Dissolve GT Detector. The Dissolve GT detector is used for the classification and declaration of dissolve transitions. The condition introduced by the Dissolve GT detector is defined in Equation 4.7, where $avgSim$ represents the average similarity value of the similarity matrix for a given dissolve GT candidate segment. If the average similarity of the first 64 frames (first batch) of the dissolve candidate segment does not satisfy the condition 4.7 then the proposed dissolve GT candidate segment is completely discarded. Otherwise, the frames of the segment are processed in batches until the average similarity value of each batch satisfies the condition 4.7. If the average similarity value of frame batch does not satisfy the condition 4.7 the frames of the batch and the rest of the proposed segment frames are discarded. All of the processed frames that satisfied the conditions form the dissolve GT.

$$\text{Dissolve classification condition: } avgSim < C_d \tag{4.7}$$

FOI / Wipe GT Detector. Similarly to the Dissolve GT detector, the FOI/Wipe GT detector is utilized for the declaration of Fade Out/In and Wipe transitions. This

detector also defines a conditional check for the average similarity value of proposed FOI / Wipe GT candidates which is provided in Equation 4.8. The decision making procedure FOI / Wipe GT declaration follows the exact flow as the Dissolve GT detector but utilizes condition 4.8 instead.

$$\text{Fade / Wipe classification condition: } avgSim < C_{fw} \tag{4.8}$$

After several empirical tests and statistical analysis, the parameters C_d and C_{fw} in the classification conditions are configured to 0.96 and 0.75 respectively. The GT detection algorithm is summarized in Algorithm 4.1.

Algorithm 4.1: GT Detection Algorithm

Input: Candidate frame pairs with a list of similarity values
Output: List of GT

```
Process
foreach frame pair do
    if 4.5 satisfied then
        Create dissolve candidate segment
    end
    if 4.6 satisfied then
        Create foi/wipe candidate segment
    end
end
foreach dissolve candidate segment do
    resultGT ← []
    foreach frame batch do
        Execute ResidualGTNetModel
        Obtain Similarity matrix
        Calculate Average Similarity Value avgSim
        if 4.7 satisfied then
            Add frames to result list resultGT
        end
        else
            Discard the segment
        end
    end
    if resultGT not empty then
        Declare frames of resultGT as a GT
    end
end
foreach foi/wipe candidate segment do
    resultGT ← []
    foreach frame batch do
        Execute ResidualGTNetModel
        Obtain Similarity matrix
        Calculate Average Similarity Value avgSim
        if 4.8 satisfied then
            Add frames to result list resultGT
        end
        else
            Discard the segment
        end
    end
    if resultGT not empty then
        Declare frames of resultGT as a GT
    end
end
return
```

CHAPTER 5

Evaluation and Results

Based on the thorough qualitative analysis of the historical films presented in Chapter 3 and the implementation described in Chapter 4, this chapter provides a detailed overview of the experiments and evaluation of the proposed SBD framework. First, the test setup and the evaluation metrics used in the experimental studies are described. Then, Section 5.3 proceeds with a quantitative analysis of historical films used for the evaluation. Section 5.4 covers all of the AT detection experiments conducted on the historical film material and the reference film material. Additionally, Section 5.4 includes a complete qualitative and comparative analysis of the AT detection performance of different configurations of the proposed framework and the state-of-the-art approaches. Finally, this chapter ends with Section 5.5 which presents and evaluates the results of the GT detection.

5.1 Experimental Setup

The proposed SBD framework, as well as all of the experiments conducted, are implemented using Python 3.6.9. The CUDA version used in this thesis is 10.1. Furthermore, for the training and testing of the CNN models, the MXNet framework with version 1.6 is used. The training of the CNN models is performed on a machine with a GPU Nvidia GeForce GTX 1080 with 8GB RAM. The CPU configuration of the machine is Intel Core i7-7820 @ 2.9GHz. The batch size used for the training of the models is configured to 32. For the image processing tasks, the standard computer vision library OpenCV is employed. The version of OpenCV utilized in this setup is 4.1.2.

5.2 Evaluation metrics

The proposed SBD framework, as well as the selected state-of-the-art approaches, are evaluated on the two historical datasets EFilms and IMC. Furthermore, the proposed approach

is validated on three benchmark datasets RAI [BGC15b], BBC Planet Earth [BGC15a] and ClipShots [TFK+18]. The performance and efficiency of the proposed SBD framework and the selected state-of-the-art approaches is analyzed and expressed in terms of the Precision, Recall, and F1-score values.

Precision is defined as the ratio of the correctly identified shot boundaries to the total number of shot boundary predictions made by a method [BR96]. Precision measures the accuracy of a method in identifying the shot boundaries. Commonly, a high precision value indicates a low number of false predictions. Equation 5.1 presents the formulation of Precision [BR96].

$$Precision = \frac{\text{Number of correctly identified shot boundaries}}{\text{Total number of identified shot boundaries}} \tag{5.1}$$

The ability of a method to identify and detect all of the shot boundaries is captured by its recall value. In contrast to precision, recall provides the ratio between the correctly identified shot boundaries by a method and the total number of actual shot boundaries present in the annotations [BR96]. A high recall value correlates to a low number of misidentifications of shot boundaries. The definition of recall is provided by Equation 5.2 [BR96].

$$Recall = \frac{\text{Number of correctly identified shot boundaries}}{\text{Total number of annotated shot boundaries}} \tag{5.2}$$

In an ideal scenario, the values of both precision and recall equal 1. In other words, this means that a method identifies all of the existing shot boundaries correctly, without making any false predictions.

The F1-score represents the harmonic mean of the precision and recall values. For this reason, the F1-score considers both the misidentified shot boundaries and the falsely identified shot boundaries. The formal definition of the F1-score is provided by Equation 5.3.

$$F1\text{-}score = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision} \tag{5.3}$$

For the evaluation of the AT detection, in this work, the standard TRECVID [SOD10] one-to-one measure is utilized. This means a prediction boundary by a method is considered a correctly identified shot boundary if at least one frame of the prediction overlaps with the annotations. For the evaluation of the GT detection, the Intersection over Union (IoU) measure is employed. The IoU measure is used to express the ability of a method to successfully detect and localize GT. It provides the ratio between the intersection and the union of the predicted GT and an annotated GT (see Equation 5.4). Hence, a GT prediction is considered correct if its IoU with the corresponding annotation exceeds a specific threshold t.

Dataset	Videos	Duration	Frames	ATs	GTs
EFilms	55	7h 40m	662715	6501	/
IMC	60	9h 25m	636724	4469	45

Table 5.1: Statistical information of the datasets EFilms and IMC.

Dataset	GTs	Dissolve	Wipe	Iris in/out	Fade in/out
IMC	45	24	5	6	11

Table 5.2: Categorization of the GTs in the IMC dataset.

$$IoU = \frac{\text{Frames identified as GT} \cap \text{Frames annotated as GT}}{\text{Frames identified as GT} \cup \text{Frames annotated as GT}} \tag{5.4}$$

5.3 Dataset Analysis

The datasets used for training of the CNN models are self-generated. In this master thesis, three training datasets are created and referred to as *HistoricalDataset*, *HistoricalDatasetGrey* and *HistoricalDataset + DA*. While the samples of the HistoricalDataset contain 3-colour channel frames, the frames of the samples of the HistoricalDatasetGrey are grayscale. Additionally, the HistoricalDataset + DA refers to the original HistoricalDataset with the added application of data augmentation techniques.

From a statistical point of view, the EFilms test dataset contains 55 historical films with a total duration of 7 hours and 40 minutes. Out of the 55 historical films, only 3 films are colour films and the rest of the 52 films are grayscale. The EFilms historical film dataset has a total number of 662715 frames and includes 6501 annotated ATs. On the other hand, the IMC test dataset consists of 60 historical films with a total duration length of 9 hours and 25 minutes. The IMC dataset contains 698624 frames and includes 4469 ATs and 45 GTs. Out of the total 45 gradual transitions, 11 transitions belong to the Fade in/out transitions category, 5 transitions belong to the Wipe transitions category, 6 transitions belong to the special group of Iris in/out transitions and the other 24 transitions represent dissolves and belong to the Dissolve transition category. Table 5.1 and Table 5.2 summarize the specifications of the two datasets.

5.4 AT Detection Experiments

This section presents the AT detection experiments conducted and discusses the results. The organization of the section is as follows. First, the experiments on the historical datasets are described and the corresponding results are analyzed and discussed. In Section 5.4.1 the experiments with the training sets of the models ATNet and ResiduaATNet

are elaborated and presented. Section 5.4.2 discusses the results of the two inference experiments and evaluates the effect of a feature combination on the AT detection performance. Next, Section 5.4.3 shows a quantitative and qualitative analysis of the false positive and negative predictions of the ResidualATNet model. The performance of the established state-of-the-art SBD approaches on the historical datasets is presented and reviewed in Section 5.4.4. Finally, the validation of the proposed AT detection approach on contemporary film material is presented in Section 5.4.5.

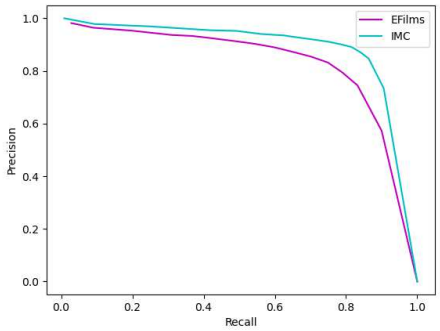
5.4.1 Training Experiments

The first set of experiments investigates the influence of the training data on the AT detection performance with respect to historical film material. The goal of the training experiments is to determine the training dataset configuration which will enable the most accurate AT detection performance in historical videos. First, the required colour space of the training dataset is determined. For this reason, the two datasets HistoricalDataset and HistoricalDatasetGrey have been created (see Section 4.2). Additionally, the effect of the training data, its size and diversity have been investigated through the application of data augmentation techniques. The data augmentation techniques applied are presented in Section 4.2.2.

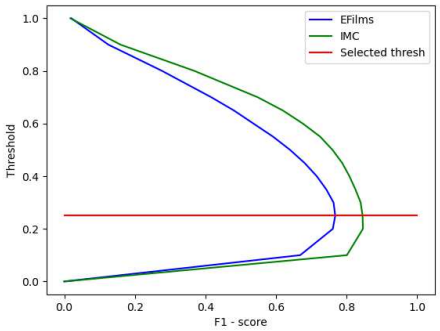
The two CNN models ATNet and ResidualATNet have been trained on the HistoricalDataset, HistoricalDatasetGrey and HistoricalDataset + DA techniques. This results in three experimental studies. Regardless of the training dataset utilized, in this set of experimental studies, the CNN models extract features from the last convolutional layer for the similarity calculation. Afterwards, in each study, every resulting model is evaluated using an adaptive and a fixed threshold. The fixed thresholds are selected with an empirical analysis of the trade-off between the precision and recall values of the models on the EFilms dataset.

Experimental Study 1 (ES1). In the first experimental study, the two CNN models ATNet and ResidualATNet are trained on the original HistoricalDataset. Figure 5.1 shows the Precision-Recall curves of ATNet and ResidualATNet on the two test datasets EFilms and IMC. Furthermore, Figure 5.1b and 5.1d capture the F1-score values achieved by the models at different thresholds. For the fixed threshold evaluation of ATNet, a value of 0.25 is selected. In contrast, for the evaluation of the ResidualATNet model, the fixed threshold equals 0.5 (see Figure 5.1).

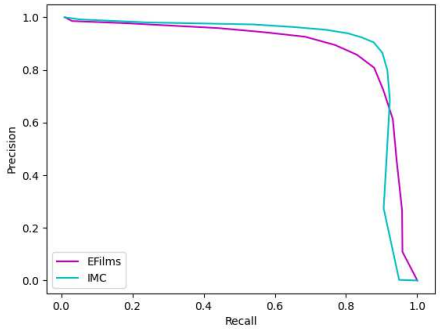
The evaluation of the models on the EFilms dataset is presented in Table 5.3. As can be observed by the results in Table 5.3, the fixed threshold evaluation of ATNet and ResidualATNet produce an F1-score of 77% and 83% respectively. Interestingly, the ATNet model also achieves an F1-score value of 77% with the adaptive threshold strategy. In contrast, the ResidualATNet model with the adaptive threshold strategy produces 81% F1-score. Even though, the adaptive threshold strategy leads to a 2% decrease in the F1-score of the ResidualATNet model it also results in a decrease in false-positive predictions from 1209 to 520.



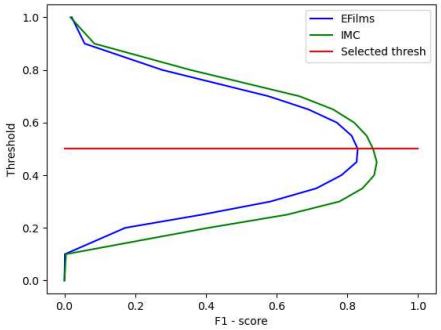
(a) Precision-Recall curves of ATNet on EFilms and IMC.



(b) F1-scores of ATNet on EFilms and IMC for different threshold values.



(c) Precision-Recall curves of ResidualATNet on EFilms and IMC.



(d) F1-scores of ResidualATNet on EFilms and IMC for different threshold values.

Figure 5.1: ES1: Precision-Recall curves and F1-score plots of the ATNet and ResidualATNet models trained on the *HistoricalDataset*.

The results achieved on the IMC dataset are presented in Table 5.4. From the results shown in Table 5.4 it can be observed that with a fixed threshold, ATNet and ResidualATNet achieve a mean F1-score of 85% and 87% respectively. Furthermore, Table 5.4 shows a 2% decrease in the F1-score of the ATNet model with an adaptive threshold. The adaptive threshold evaluation of ATNet also shows an increase in the false positive and false negative predictions from 841 and 564 with a fixed threshold to 923 and 586 with an adaptive threshold. In the case of ResidualATNet, the adaptive threshold strategy neither decreases nor increases the F1-score value which equals 87%. However, the adaptive threshold evaluation of ResidualATNet significantly reduces the number of false-positive detections from 982 with a fixed threshold to 405. Overall, the results in Table 5.3 and Table 5.4, show that with an adaptive threshold the ResiudalATNet architecture outperforms the ATNet architecture by 4% on both test datasets. With a fixed threshold, ResidualATNet still outperforms ATNet by 6% on the EFilms dataset and by 2% on the IMC dataset.

Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet	Fix	0.77	0.79	0.77	5117	1381	1384
ATNet	Adaptive	0.75	0.83	0.77	5175	1579	1143
ResidualATNet	**Fix**	**0.83**	**0.86**	**0.83**	**5618**	**1209**	**883**
ResidualATNet	Adaptive	0.89	0.77	0.81	4925	520	1576

Table 5.3: ES1: Results on the EFilms dataset achieved by the models trained on HistoricalDataset.

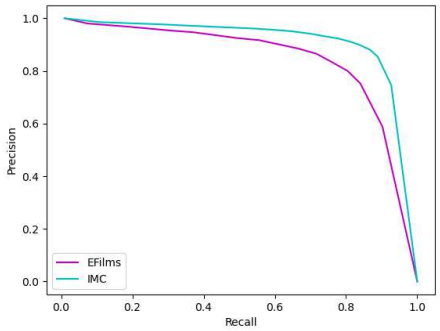
Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet	Fix	0.84	0.87	0.85	3861	841	564
ATNet	Adaptive	0.81	0.88	0.83	3839	923	586
ResidualATNet	Fix	0.84	0.92	0.87	4100	982	369
ResidualATNet	**Adaptive**	**0.90**	**0.85**	**0.87**	**3649**	**405**	**820**

Table 5.4: ES1: Results on the IMC dataset achieved by the models trained on HistoricalDataset.

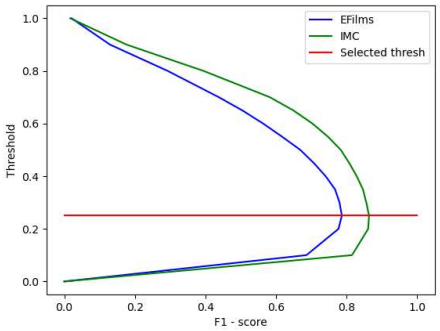
Experimental Study 2 (ES2). The two CNN models ATNet and ResidualATNet, in the second study, are trained on the HistoricalDatasetGrey. The Precision-Recall trade-off curves of the models on the test datasets are shown in Figure 5.2. For the fixed threshold evaluation in this study, based on the plots shown in Figure 5.2b and 5.2d, the thresholds are configured to 0.25 and 0.4 for ATNet and ResidualATNet, respectively. Table 5.5 and Table 5.6 provide the evaluation results of the performance of the models on the EFilms and the IMC dataset.

In contrast to the results achieved in ES1, an initial examination of the results obtained in this study shows a 2% increase in the F1-score performance of the ATNet model on both test datasets. However, this is not the case for the ResidualATNet model whose F1-score performance is not significantly different than the one achieved in ES1. As can be seen by Table 5.5, on the EFilms dataset, ATNet and ResidualATNet with a fixed threshold achieve an F1-score of 79% and 82% respectively. In comparison to ES1, the fixed evaluation of ATNet makes 110 less false positive and 115 less false negative predictions whereas the fixed evaluation of ResidualATNet in this study makes 372 more false-negative predictions and 454 less false positive predictions. With an adaptive threshold, both models experience a 1% drop in their F1-score value which equals 78% for ATNet and 81% for ResidualATNet. On the other hand, on the IMC dataset ATNet and ResidualATNet with a fixed threshold produce F1-score values of 86% and 88%, respectively. Similarly to the performance on the EFilms, the adaptive threshold evaluation of the models is characterised by lower F1-score values of 85% and 86% (see Table 5.6).

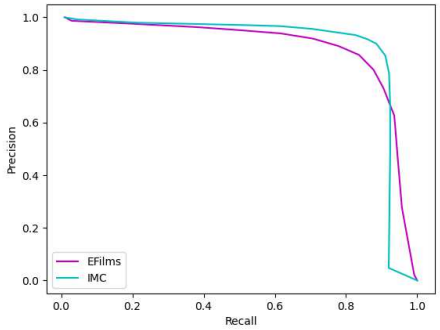
The results in this study show that the conversion of the training set from 3 colour



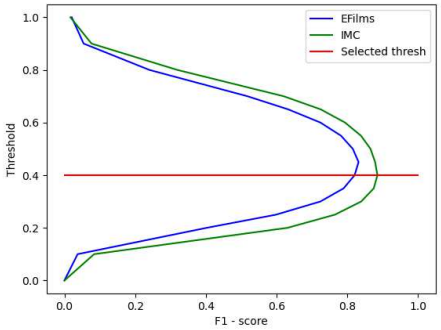
(a) Precision-Recall curves of ATNet on EFilms and IMC.



(b) F1-scores of ATNet on EFilms and IMC for different threshold values.



(c) Precision-Recall curves of ResidualATNet on EFilms and IMC.



(d) F1-scores of ResidualATNet on EFilms and IMC for different threshold values.

Figure 5.2: ES2: Precision-Recall curves and F1-score plots of the ATNet and ResidualATNet models trained on the *HistoricalDatasetGrey*.

channels to 1 colour channel i.e. grey space leads to a 2% increase in the F1-score of ATNet and 1% decrease in the F1-score of ResidualATNet. Despite the increase in performance of the ATNet model on EFilms and IMC, the ResidualATNet model produces superior results with both threshold evaluation policies. Consequently, it can be concluded that for the self-generated HistoricalDataset and the two models ATNet and ResidualATNet, the colourspace conversion to HistoricalDatasetGrey does not have a significant impact on the transition detection performance of the two models. This could be due to the fact that most of the samples of the HistoricalDataset were already in the greyscale colour region.

Experimental Study 3. In the last experimental study, ATNet and ResidualATNet are trained on the original (RGB) HistoricalDataset with the additional application of data augmentation techniques. The data augmentation techniques are applied "on-the-fly" during training. The fixed thresholds in this study are selected in an equivalent manner

Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet	Fix	0.80	0.80	0.79	5232	1272	1269
ATNet	Adaptive	0.76	0.83	0.78	5309	1480	1192
ResidualATNet	**Fix**	**0.88**	**0.80**	**0.82**	**5246**	**755**	**1255**
ResidualATNet	Adaptive	0.88	0.78	0.81	4981	555	1520

Table 5.5: ES2: Results obtained on the EFilms dataset achieved by the models trained on the HistoricalDatasetGrey.

Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet	Fix	0.87	0.88	0.86	3898	747	571
ATNet	Adaptive	0.83	0.89	0.85	3862	859	607
ResidualATNet	**Fix**	**0.89**	**0.90**	**0.88**	**3972**	**589**	**497**
ResidualATNet	Adaptive	0.90	0.85	0.86	3681	412	788

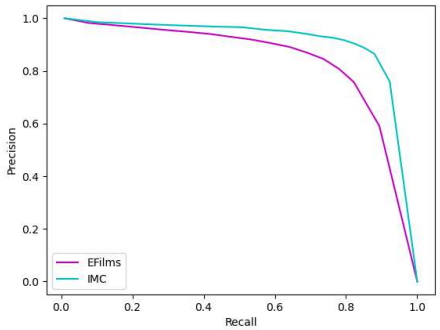
Table 5.6: ES2: Results on the IMC dataset achieved by the models trained on the HistoricalDatasetGrey.

to ES1 and ES2 and which is based on the empirical examination of the Precision-Recall trade-off curves shown in Figure 5.3. The fixed threshold of the ATNet and ResidualATNet models is configured to 0.3 and 0.4, respectively (see Figure 5.3).

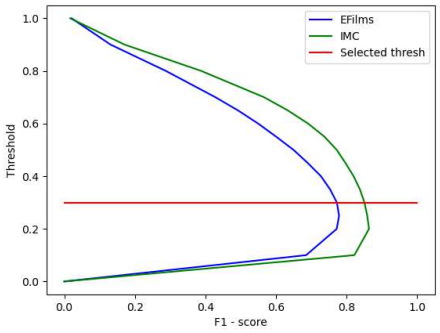
The effects of the data augmentation techniques and efficiency of the models on the EFIlms and IMC datasets are captured by the results shown in Table 5.7 and Table 5.8. On the EFilms dataset, with a fixed threshold ATNet and ResidualATNet achieve F1-score values of 77% and 82% respectively. For both models, the fixed threshold evaluation is superior to the adaptive by 1% F1-score. In this experimental study, on the EFilms dataset, the fixed threshold evaluation ATNet makes 2022 false-positive predictions. This is a significant increase from the 841 false-positive predictions in ES1 and the 1272 false-positive predictions in ES2.

In contrast to the EFilms dataset, on the IMC dataset, the ATNet model achieves an F1-score value of 85% with both threshold policies. Furthermore, the ResidualATNet model achieves an F1-score of 88% with a fixed threshold and 86% with an adaptive threshold. According to the results shown in Table 5.7 and Table 5.8 the application of flips does improve the detection performance of the ATNet model. In the case of ATNet, the data augmentation techniques lead to an increase in false-positive transition detections. This is not the case in the results of fixed threshold evaluation of ResidualATNet in which a decrease in the false positive predictions is observed. The adaptive evaluation of ResidualATNet, on the other hand, does not show significant improvement to the results achieved in ES1.

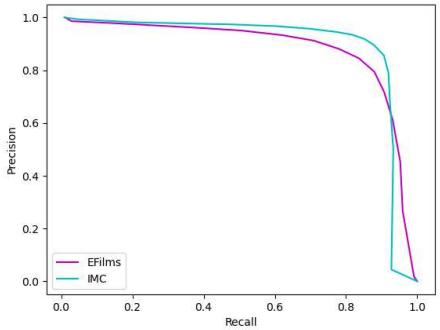
Discussion. The charts in Figure 5.4 and Figure 5.5 summarize the results of all of



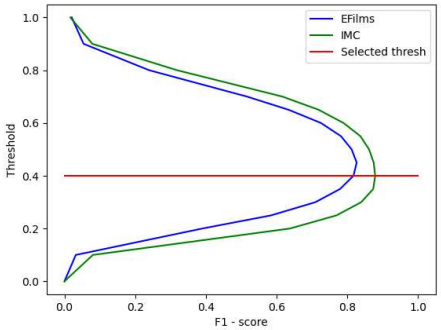
(a) Precision-Recall curves of ATNet on EFilms and IMC.



(b) F1-scores of ATNet on EFilms and IMC for different threshold values.



(c) Precision-Recall curves of ResidualATNet on EFilms and IMC.



(d) F1-scores of ResidualATNet on EFilms and IMC for different threshold values.

Figure 5.3: ES3:Precision-Recall curves and F1-score plots of the ATNet and ResidualATNet models trained on the *HistoricalDataset + DA*.

Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet	Fix	0.74	0.85	0.77	5527	2022	974
ATNet	Adaptive	0.74	0.82	0.76	5304	1672	1197
ResidualATNet	**Fix**	**0.88**	**0.79**	**0.82**	**5212**	**727**	**1289**
ResidualATNet	Adaptive	0.88	0.78	0.81	4982	553	1519

Table 5.7: ES3: Results on the EFilms dataset achieved by the models trained on the HistoricalDataset + DA.

Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet	Fix	0.83	0.90	0.85	4005	1147	464
ATNet	Adaptive	0.82	0.90	0.85	3881	976	588
ResidualATNet	**Fix**	**0.88**	**0.89**	**0.88**	**3959**	**615**	**510**
ResidualATNet	Adaptive	0.89	0.85	0.86	3681	436	788

Table 5.8: ES3: Results on the IMC dataset achieved by the models trained on the HistoricalDataset + DA.

the training experiments on the two historical datasets EFilms and IMC respectively. From the results presented in the charts, it can be concluded that the models trained on HistoricalDataset and HistoricalDatasetGrey produce similar results. Furthermore, there is no significant difference between the results achieved by the models trained on the HistoricalDataset + DA and the models trained without the application of data augmentation techniques. One reason for this could be the fact that flips neither increase nor decrease the overall performance of the models on historical data.

A crucial aspect captured by the charts in Figure 5.4 and Figure 5.5 is the superiority of the ResidualATNet model to the ATNet model. On the EFilms dataset, the best F1-score of 83% is achieved by the ResidualATNet model trained on the HistoricalDataset with a fixed threshold. On the IMC dataset, the best F1-score of 88% is produced by the ResidualATNet model trained on the HistoricalDatasetGrey with a fixed threshold. Consequently, the CNN architecture of ResidualATNet outperforms the CNN architecture of ATNet on both datasets.

Finally, the charts demonstrate that the utilization of an adaptive threshold policy leads to competitive results. According to the results, the adaptive threshold evaluation achieves the same F1-score values as the fixed threshold evaluation in the best-case scenario. In the worst-case scenario, the adaptive threshold evaluation shows a 2% decrease in the mean F1-score. Nonetheless, the adaptive threshold evaluation policy results in robust and generalizable models and reduces the number of false-positive transitions in the case of ResidualATNet.

5.4.2 Inference Experiments

This section presents the experiments conducted during the inference phase of the models ATNet and ResidualATNet. The inference experiments aim to improve the results and performance of the models achieved in the training experiments. In the training experiments, the models trained on the original HistoricalDataset and HistoricalDatasetGrey produced similar results with the highest F1-score values (see Section 5.4.1). Therefore, for the inference experiments, the models of ATNet and ResidualATNet trained on the original HistoricalDataset and HistoricalDatasetGrey are taken under consideration.

The inference experiments focus on examining the contribution of complementary infor-

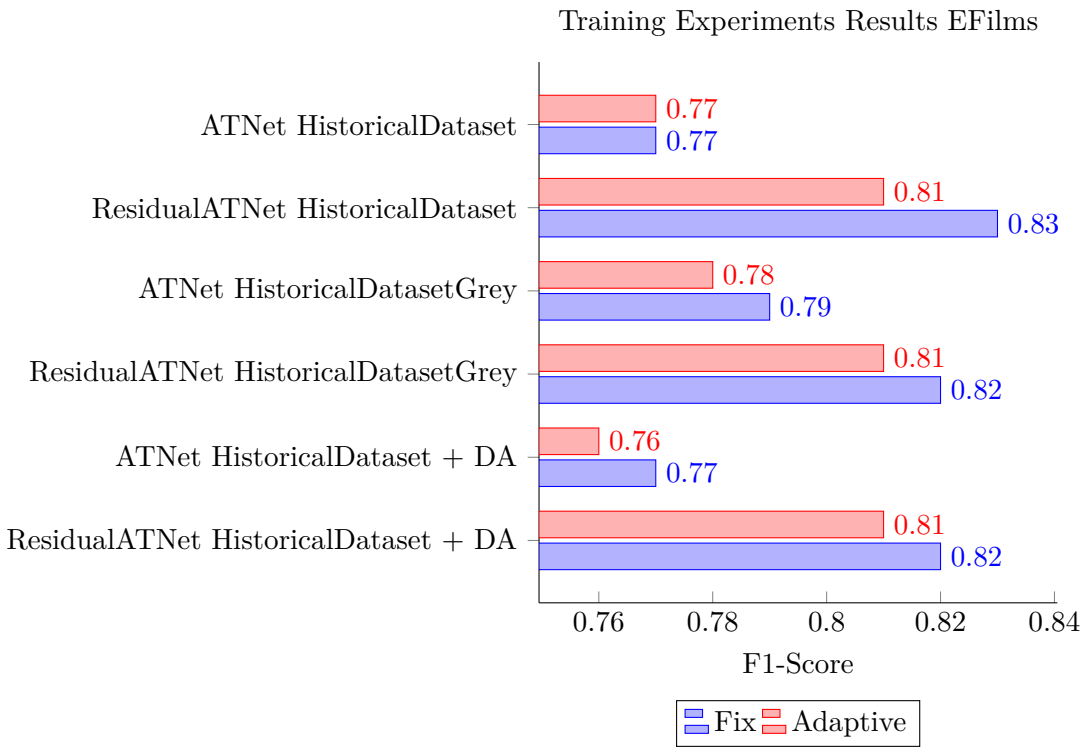


Figure 5.4: Complete results of the Training experiments on the EFilms dataset.

mation to the SBD process in historical videos. In this master thesis, complementary information is provided during the inference process in two ways: by extracting frame information from a combination of multiple features and by defining target areas within a frame and extracting different information from each area. This results in two experimental studies carried out with the two CNN architecture models ATNet and ResidualATNet.

Experimental Study 4 (ES4). This experimental study investigates the effect of extracting multi-feature information on the AT detection process. Instead of relying on single-feature information, in this study, the two CNN architecture models ATNet and ResidualATNet extract frame information from three different features. ATNet extracts the convolutional features from the 7th, 10th and 13th convolutional layer. Moreover, ResidualATNet utilizes the feature information from the 7th, 11th and 15th convolutional layer. Similarly to training experiments, the evaluation of ATNet and ResidualATNet is carried out with an adaptive as well as a fixed threshold. Furthermore, the fixed threshold values in this study are selected empirically and are configured to 0.55 and 0.6 for ATNet and ResidualATNet, respectively (see Figure 5.6). The results of this experimental study are summarized in Table 5.9 and Table 5.10.

TTable 5.9 presents the performance on the EFilms dataset achieved by the models ATNet and ResidualATNet trained on HistoricalDataset and HistoricalDatasetGrey. As can be observed in Table 5.9, the utilization of multiple feature information has an

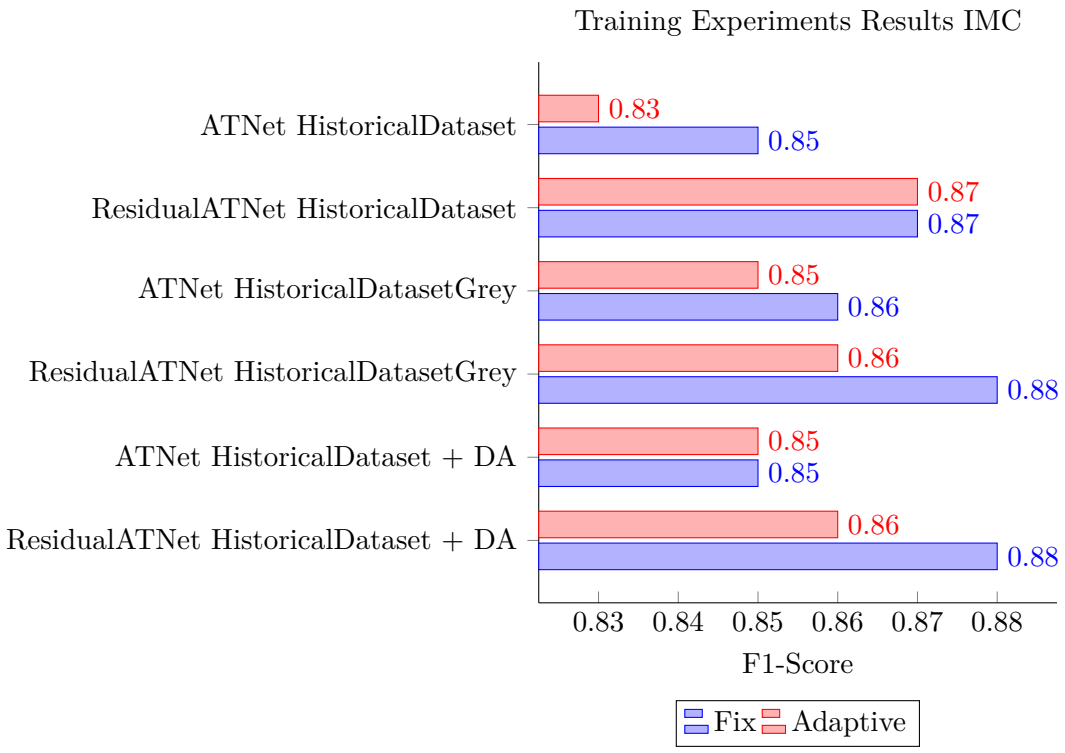
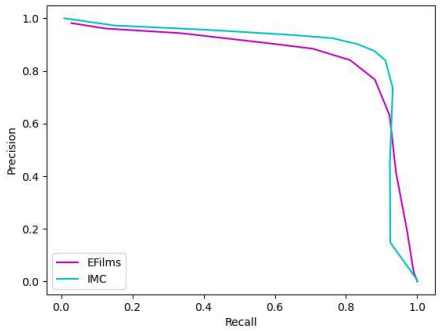


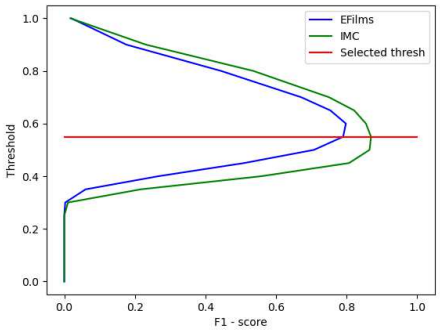
Figure 5.5: Complete results of the Training experiments on the IMC dataset.

overall positive effect on the AT detection performance in the historical film material. Specifically, it helps ATNet and ResidualATNet trained on HistoricalDataset achieve an F1-score of 79% and 84% with a fixed and 82% and 83% with an adaptive threshold. When trained on HistoricalDatasetGrey, the models ATNet and ResidualATNet achieve an F1-score of 81% and 84% with a fixed and 82% and 83% with an adaptive threshold. Although the utilization of the greyscale dataset improves the F1-score of ATNet with a fixed threshold by 2%, it does not affect the F1-score values of ResidualATNet. In fact, ResidualATNet with a fixed threshold predicts 113 transitions more when trained on the HistoricalDataset than when using HistoricalDatasetGrey.

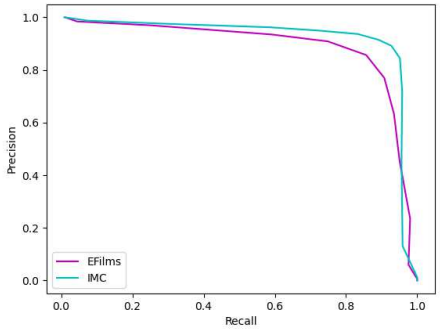
Another point illustrated by the results in Table 5.9 is that the utilization of multi-feature information paired with an adaptive threshold policy results in high precision values. With the HistoricalDataset, ResidualATNet achieves a precision value of 91%. This result is noteworthy, given the special properties of historical films and the additional difficulties they pose to the problem of SBD. The high precision value is due to the adaptive threshold utilization which results in a sharp decrease in the number of false transition identifications. ResidualATNet with an adaptive threshold makes only 330 false-positive detections which is 3 times less than when compared to the fixed threshold evaluation. However, the adaptive threshold policy also produces a slight increase in the number of non-detected transitions, hence the lower recall value of 78%. This evident



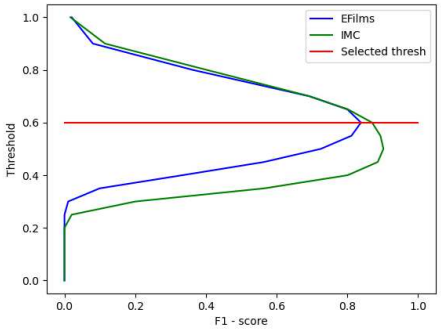
(a) Precision-Recall curves of ATNet-3f on EFilms and IMC.



(b) F1-scores of ATNet-3f on EFilms and IMC for different threshold values.



(c) Precision-Recall curves of ResidualATNet-3f on EFilms and IMC.



(d) F1-scores of ResidualATNet-3f on EFilms and IMC for different threshold values.

Figure 5.6: ES4: Precision-Recall curves and F1-score plots of the ATNet-3f and ResidualATNet-3f models trained on the *HistoricalDataset*.

trade-off between precision and recall values is captured by Table 5.9.

The results on the IMC dataset are presented in Table 5.10. The ResidualATNet model trained on the HistoricalDataset and evaluated with an adaptive threshold reaches the highest F1-score of 90%. When comparing the adaptive threshold evaluations, ResidualATNet outperforms ATNet in terms of both precision and recall values. The ResidualATNet model successfully detects 3805 ATs. Furthermore, according to the results in Table 5.10, the ResidualATNet model produces only 237 false-positive detections. This means, in contrast to ATNet, ResidualATNet is more successful in combating the artefacts and obstacles of historical film material which makes the model adequate for performing SBD in historical films.

Discussion. The chart presented in Figure 5.7 demonstrates a comparison of the results achieved in ES1 against the results achieved in ES4. The chart compares the

Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet-3f	Fix	0.86	0.77	0.79	4977	728	1524
ATNet-3f	Adaptive	0.83	0.84	0.82	5206	878	1112
ResidualATNet-3f	**Fix**	**0.86**	**0.86**	**0.84**	**5706**	**973**	**795**
ResidualATNet-3f	Adaptive	0.91	0.78	0.83	5009	330	1492
ATNet-3f-grey	Fix	0.88	0.78	0.81	5153	735	1348
ATNet-3f-grey	Adaptive	0.84	0.83	0.82	5339	854	1162
ResidualATNet-3f-grey	Fix	0.87	0.84	0.84	5593	778	908
ResidualATNet-3f-grey	Adaptive	0.92	0.78	0.83	5024	323	1477

Table 5.9: ES4: Results on the EFilms dataset using a multi-feature extraction strategy achieved by the models trained on the *HistoricalDataset* and *HistoricalDatasetGrey*.

Model	Thresh	Precision	Recall	F1-score	TP	FP	FN
ATNet-3f	Fix	0.88	0.88	0.87	3929	578	496
ATNet-3f	Adaptive	0.86	0.88	0.86	3826	579	599
ResidualATNet-3f	Fix	0.83	0.94	0.87	4187	1386	282
ResidualATNet-3f	**Adaptive**	**0.93**	**0.88**	**0.90**	**3805**	**237**	**664**
ATNet-3f-grey	Fix	0.90	0.89	0.89	3974	551	495
ATNet-3f-grey	Adaptive	0.88	0.89	0.88	3857	572	612
ResidualATNet-3f-grey	Fix	0.85	0.93	0.88	4171	1130	298
ResidualATNet-3f-grey	Adaptive	0.94	0.87	0.89	3779	226	690

Table 5.10: ES4: Results on the IMC dataset using a multi-feature extraction strategy achieved by the models trained on the *HistoricalDataset* and *HistoricalDatasetGrey*.

performance of the models ATNet and ResidualATNet using single feature information to ATNet and ResidualATNet using multi-feature information. The chart in Figure 5.7 presents the performance of the models trained on the HistoricalDataset and includes only their adaptive threshold evaluation. The performance of ATNet and ResidualATNet is represented in terms of the F1-score values achieved. From the results presented in Figure 5.7, it can be inferred that the utilization of multiple feature information leads to overall higher F1-score values for ATNet and ResidualATNet on both historical datasets. In comparison to the single feature information, the combination of multiple features results increases both precision and recall values. This means the combination of multiple features significantly enhances the ability of the models to correctly identify AT transitions against non-transitions in the historical film material. Therefore, the performance of ATNet and ResidualATNet which extract multiple feature information is characterized by a low number of false positives as well as false negatives. Overall, in the context of SBD in the historical film domain, the combination of multiple features yields great performance benefits and is therefore preferred over the dependence on single feature information.

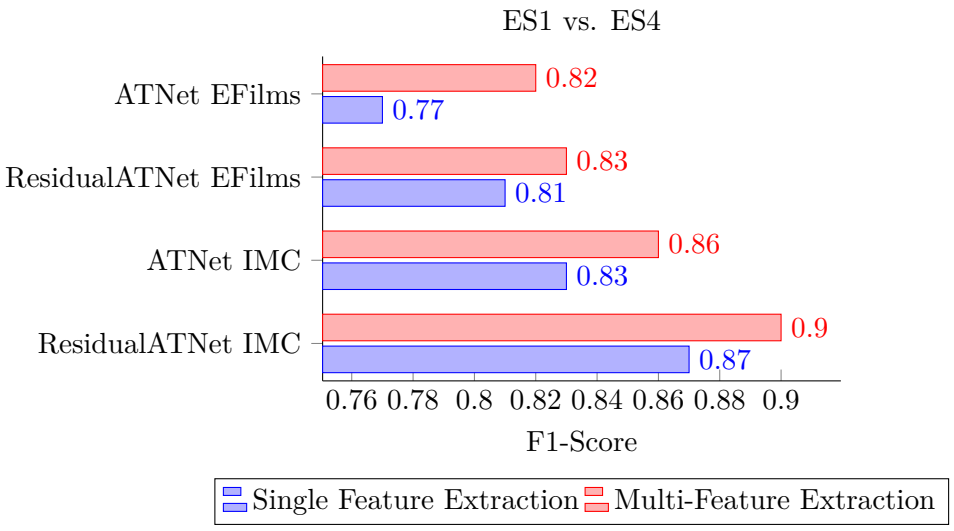
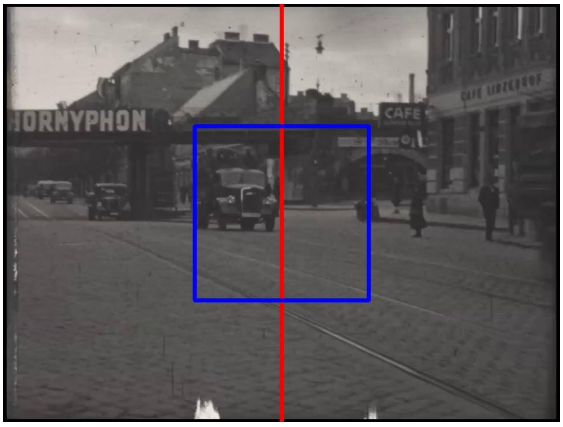


Figure 5.7: Comparison of the results achieved with the two feature extraction strategies: single feature extraction vs. multi-feature extraction

Experimental Study 5 (ES5). This experimental study aims to further improve the AT detection performance of the multi-feature extraction architecture of ATNet and ResidualATNet from ES4. Therefore in this study, only the models trained on the HistoricalDataset are considered. In this experimental study, each frame is split into three tiles before being forwarded to the CNN models. First, the frame is split vertically into two equal parts and resized to 300x300 pixels. The third tile is generated by creating a 300x300 pixels centre crop of the original frame. The visual separation of a frame into three tiles is presented in Figure 5.8. Each tile is individually processed by the CNN model which extracts a combination of features to compute the similarity value. The similarity value is compared against a fixed threshold. The fixed thresholds are selected by observing the Precision-Recall trade-off curves of the models presented in Figure 5.9. Based on the plots shown in Figure 5.9 the thresholds are configured to 0.65 for ATNet and 0.6 for ResidualATNet. The classification of an AT is performed using a voting algorithm which depending on the number of votes declares either an AT or non-AT.

Table 5.11 outlines the results achieved by ATNet and ResidualATNet on the historical datasets during this study. As demonstrated by the results in Table 5.11, ResidualATNet outperforms ATNet on both historical datasets. In this study, ATNet and ResidualATNet produce F1-score values of 74% and 79% on the EFilms dataset. Moreover, on the IMC dataset, ATNet and ResidualATNet achieve F1-scores of 84% and 85% respectively. According to the results in Table 5.11, splitting the frame into tiles does not have a positive impact on the AT detection performance of the models with respect to historical films. One reason for this is the fact that historical films are challenging and contain complex artefacts (see Section 3.3). As discussed in Section 3.3, many of the artefacts found in historical films are local. This means the artefacts only affect a small region



(a) Original frame.

(b) Left tile.

(c) Center tile.

(d) Right tile.

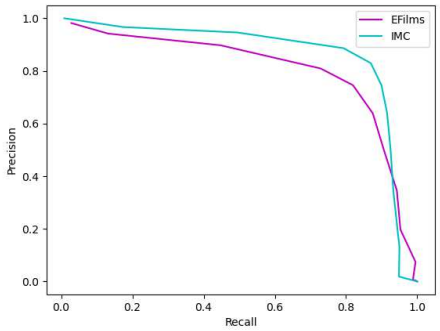
Figure 5.8: ES5: Splitting of a frame into tiles.

Model	Thresh	Test Dataset	Precision	Recall	F1-score	TP	FP	FN
ATNet-3f	Fix	EFilms	0.82	0.75	0.74	4832	1421	1669
ResidualATNet-3f	Fix	EFilms	0.85	0.76	0.79	5143	1001	1358
ATNet-3f	Fix	IMC	0.87	0.83	0.84	3567	737	902
ResidualATNet-3f	Fix	IMC	0.84	0.90	0.85	3996	1314	473

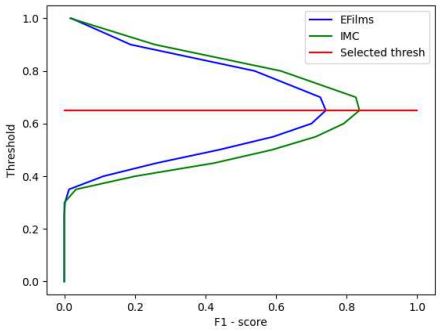
Table 5.11: ES5: Results on the EFilms and IMC dataset achieved by separating the frames into tiles.

of the frame. By considering the information of the tiles (parts) of the frames, this method is very sensitive to local changes in the frames. Consequently, when the tiles of subsequent frames are compared, the artefacts result in low similarity values and ultimately yield a high number of false-positive detections. This could be due to the insufficient number of tiles extracted from each frame.

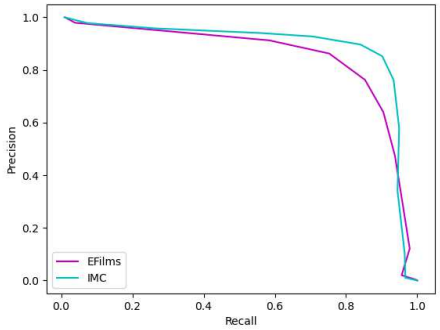
Discussion. The chart in Figure 5.10 depicts a comparison of the performance of the multi-feature extraction models of ATNet and ResidualATNet which use the full frames



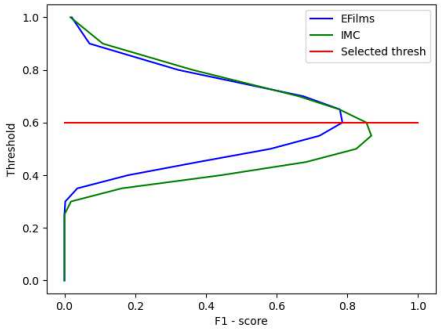
(a) Precision-Recall curves of ATNet on EFilms and IMC.



(b) F1-scores of ATNet on EFilms and IMC for different threshold values.



(c) Precision-Recall curves of ResidualATNet on EFilms and IMC.



(d) F1-scores of ResidualATNet on EFilms and IMC for different threshold values.

Figure 5.9: ES5: Precision-Recall curves and F1-score plots of the ATNet-3f and ResidualATNet-3f using splitted frames into tiles.

as input (ES4) against the multi-feature extraction models that use the frame tiles as input (ES5). The models presented in the chart are trained on the HistoricalDataset. The performance of the models is expressed in terms of the F1-score achieved. Moreover, performance values for both datasets EFilms and IMC are provided. Overall, splitting the frames into tiles does not increase the AT detection performance in the historical film material, as evident in Figure 5.10. The separation into tiles in combination with the multi-feature information extraction makes the CNN models extremely sensitive to small content changes. The sensitivity of the CNN models to small content changes becomes a major disadvantage in the context of historical videos. The low quality of the historical videos, their local artefacts as well as the instability and shaking of the films all lead to many false-positive AT predictions. Therefore, when using a combination of feature information, it is advantageous to utilize the full-frame instead of parts i.e. tiles of the frame. The utilization of the full frames as in ES4 provides a stable similarity measure that is less sensitive to the artefacts found in historical films.

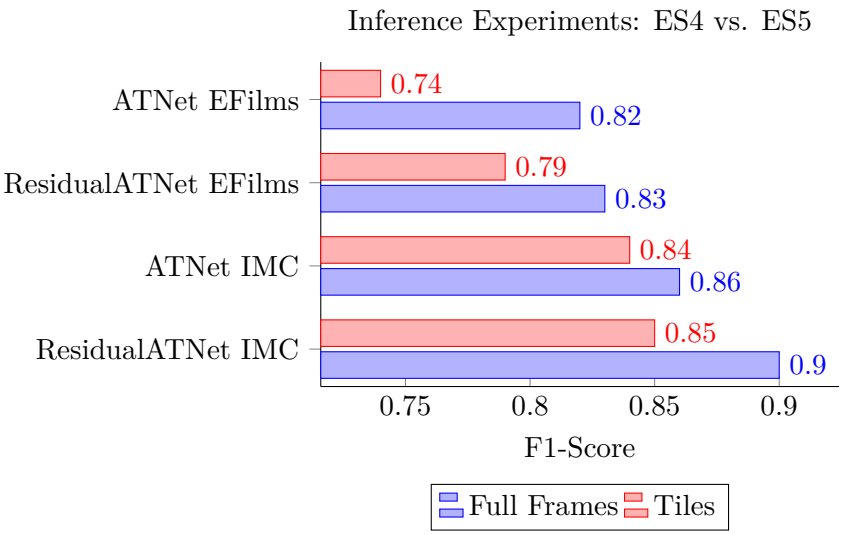


Figure 5.10: Comparison of the results achieved in the two inference experiments: full-frame processing vs. splitting the frames into tiles.

5.4.3 Qualitative and Quantitative Analysis

The AT detection analysis is performed on the results obtained by the CNN architecture ResidualATNet-3f. The ResidualATNet-3f model under consideration utilizes a combination of multiple features and an adaptive threshold policy. The results analyzed in this section are achieved and presented in the ES4. The analysis of the AT detection predictions of the ResidualATNet-3f model is performed manually. The analysis of the ResidualATNet-3f model involves an examination of the inference speed of the model and contains a detailed investigation of the annotated transitions and the false positive predictions of the model. Finally, this section also includes a thorough analysis of the transitions that the ResidualATNet-3f model fails to identify (i.e. the false-negative cases).

Inference Speed

The inference speed of the ResidualATNet-3f model is evaluated and analyzed on two different machines. The first machine is configured with a Tesla P100 PCIe GPU with 16GB memory and an Intel Xeon CPU with 2.30GHz. The second machine is equipped with an Nvidia GeForce GTX 1080 GPU with 8GB memory and an Intel Core i7-7820 with 2.9GHz. The graphs depicted in Figure 5.11 represent the trendlines of the inference of speed of the ResidualATNet-3f model for a given input frame size on GPU as well as on CPU.

On the Tesla P100 GPU, the ResidualATNet-3f model can process 512 input frames in 1.13 seconds. Furthermore, on the aforementioned GPU, the model can process 256

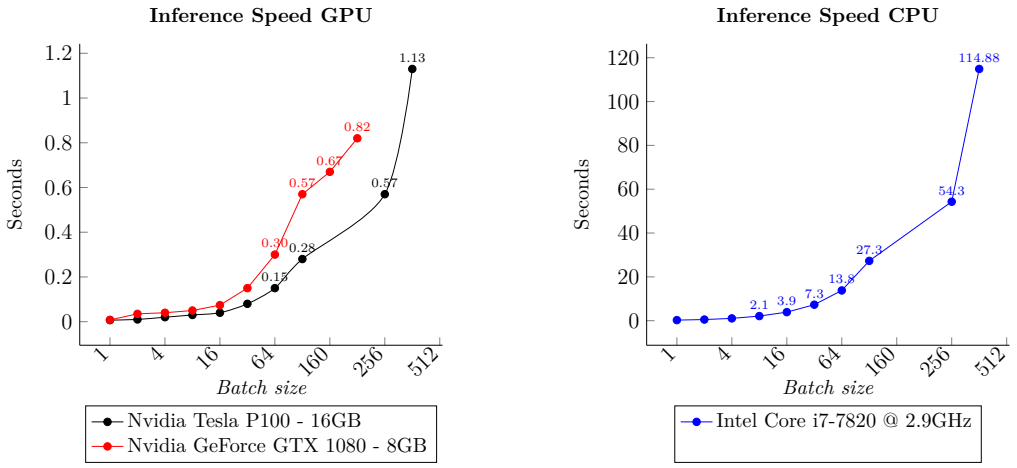


Figure 5.11: Comparison of the inference speed achieved by ResidualATNet-3f on GPU and CPU.

frames in 0.57 seconds and 128 frames in 0.28 seconds. For a historical film which runs at 16 FPS, the ResidualATNet-3f model obtains 28 x real-time or 454 FPS.

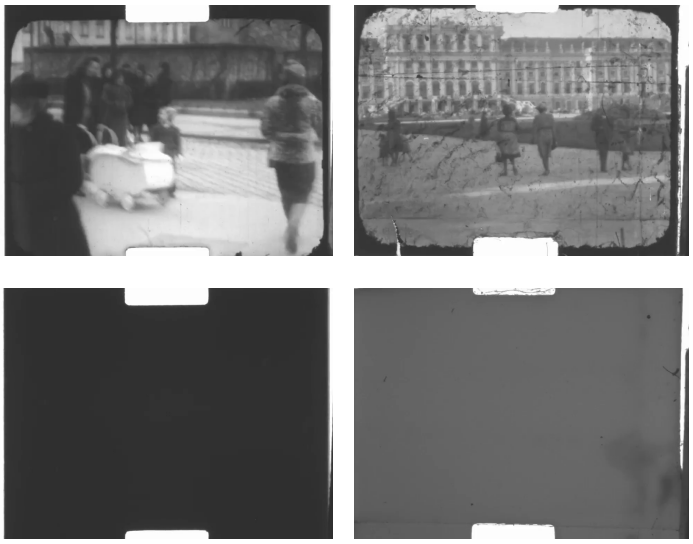
In contrast, on the machine configured with GeForce GTX 1080 GPU, the ResidualATNet-3f model can process 128 frames in 0.57 seconds. Moreover, the computational power and memory of GeForce GTX 1080 GPU allow a maximum input size of 192 frames which can be processed in 0.82 seconds. In this case, the ResidualATNet-3f model obtains 15 x real-time speed up or 234 FPS for a historical film of 16 FPS. Finally, the performance of ResidualATNet-3f model is also evaluated on the Intel Core i7-7820 with 2.9GHz processor. In this setup, the model can process 128 frames in 27.3 seconds.

Annotations and False Positives

The false-positive analysis reveals two crucial facts. First, the two historical datasets contain a vast number of ATs which are not included in the annotations. This means that the model successfully detects and localizes ATs in the historical films, which are not labelled in the annotations and are therefore incorrectly considered as false positives. The ATs which are not included in the annotations involve cases of similar scenes, random interruptions of the frame sequence with a black, grey or a white frame as well as the start and the end positions of intertitle frame sequences. Second, the actual false positive cases where the model falsely identifies transitions are due to the artefacts found in historical films. In this case, the most common contributors to the false positive detections are blurred frames, scratches, tears, colour fades and low contrast. The false-positive analysis is performed separately for the two historical datasets.

EFilms Dataset. On the EFilms dataset, the ResidualATNet-3f model makes a total of 330 false-positive predictions. A thorough manual examination of the 330 false-positive cases shows that 75 of them refer to an actual transition. This means that ca. 23% of

the false-positive cases in the EFilms dataset are incorrectly considered false positive as they represent a true AT. Moreover, this also shows that 75 transitions labels are not included in the EFilms dataset annotations. These 75 incorrectly labelled transitions contained many cases of sudden interruptions of the frame sequence with a monochrome frame such as black, grey and white frames. Figure 5.12 illustrates actual examples of such cases. Furthermore, a small subset of these 75 transitions represents frames with very similar scenes which in reality are separated by an AT. A visual illustration of three such transitions whose label is not included in the annotations is presented in Figure 5.13. By updating the number of false and true positive predictions, the overall precision value of the ResidualATNet-3f model increases by 1%.



(a) AT with a black frame. (b) AT with a grey frame.

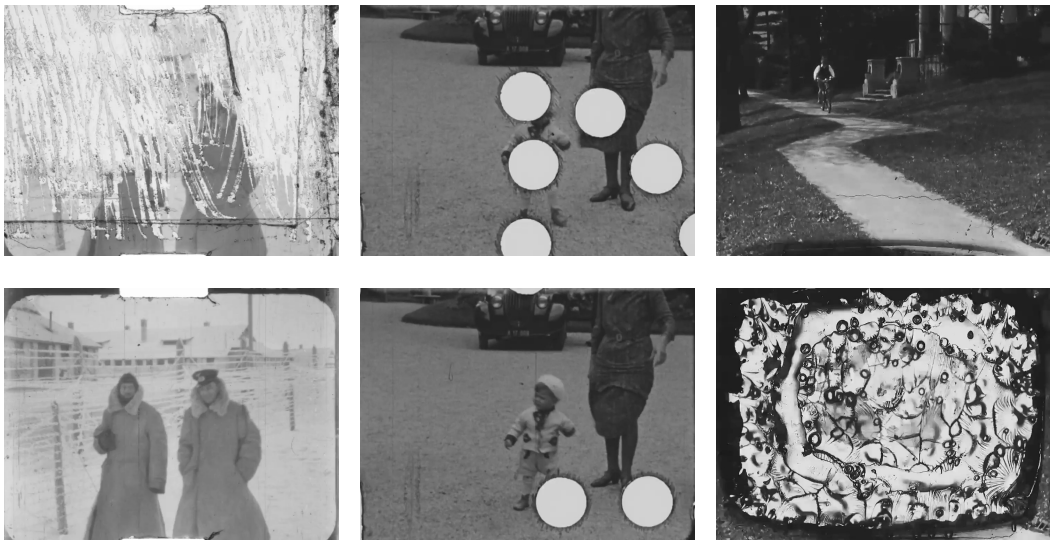
Figure 5.12: Detected sudden content interruptions with monochrome frames.

Nevertheless, the rest of the 255 predictions represent actual false-positive cases. Interestingly, the manual analysis shows that the false positive predictions mostly refer to frames which are filled with artefacts. The false-positive predictions involve very damaged frames with scratches, cue dots, spills and mould. Figure 5.14 illustrates such false positive predictions. Moreover, there are cases where local damage is so great that only a fraction of the frame is visible (see Figure 5.15a). The ResidualATNet-3f model also reports transitions between frames where the top of the frame represents the bottom half of the previous frame and the bottom of the frame represents the top of the next frame (see Figure 5.15b). This artefact is known as a filmstrip tear, where the ends of a filmstrip are incorrectly reattached and the tear remains visible. Another major problem in the actual false positives is the lack of colour i.e. low contrast of the frames (see Figure 5.15c). As discussed in section 3.3 low contrast is the result of the incorrect storage and repeated playback and replication of the historical films. Finally, a large portion of the false positive predictions involved blurred frames or frames of very low

quality in general. These two artefacts are the byproduct of the instability or shaking in historical films which are magnified through the copying of the filmstrips. Examples of the false positive predictions reported by the ResidualATNet-3f model are demonstrated in Figure 5.15.

(a) (b) (c)

Figure 5.13: Detected boundaries not included in the EFilms ground truth annotations.



(a) Scratches (b) Cue dots (c) Chemical damage

Figure 5.14: False positive predictions by the ResidualATNet-3f model due to the artefacts in the EFilms dataset.

(a) Scratch (b) Tear (c) Low contrast (d) Blur

Figure 5.15: False positive detection due to tears, scratches, blur and colour fades.

IMC Dataset. On the IMC dataset, the ResidualATNet-3f model produces a total number of 237 false-positive predictions. Since the IMC dataset contains a vast number of intertitle frames, the start and endpoints of the intertitle frame sequences form the majority of the false-positive cases. Even though the start and end positions of intertitle frame sequences are included in the annotations, they are incorrectly labelled as GT. As a result, actual ATs such as the examples shown in Figure 5.16a and 5.16b are considered false-positive predictions. Out of the 237 false-positive predictions, 121 (84%) cases represent ATs not captured by the annotations of the IMC dataset. Apart from the start and end positions of an intertitle frame sequence, some of the incorrectly considered false-positive predictions refer to standard transitions whose label is omitted in the annotations. Figure 5.16c and 5.16d demonstrate transition examples whose label is not present in the IMC dataset. The rest of the 116 actual false positive predictions involve frames affected by artefacts. Similarly to the false positive predictions on the EFilms dataset, these cases correspond to heavily blurred and low contrast frames, frames spills, mould as well as frames filled with scratches. Figure 5.17a and Figure 5.17b demonstrate examples of the actual false positive predictions of the ResidualATNet-3f model on the IMC dataset. Using the corrected annotations which consider the 121 cases as true positives, the evaluation metrics on the IMC dataset significantly improve. With the corrected annotations, the F1-score of the ResidualATNet-3f model increases by 1%.

Taking these learned facts from the manual false-positive analysis into account, Table 5.12 contains the updated evaluation of the ResidualATNet-3f model on the two historical datasets EFilms and IMC with corrected labels. As can be observed by the results in Table 5.12, the precision values on the two datasets exceed 95%. Furthermore, the recall values equal 77% and 86% on the EFilms dataset and IMC dataset respectively. Finally, the ResidualATNet-3f model achieves an overall F1-score of 85% on the EFilms dataset and 91% on the IMC dataset. This performance of ResidualATNet-3f is remarkable considering the complexity and challenges of historical films.

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

Figure 5.16: Examples of not annotated intertitle-frames (a) - (b) and transitions (c) - (d) in the IMC dataset.

(a) Colour fade.

(b) Blur, shaking and wrong exposure.

Figure 5.17: False positive predictions of the ResidualATNet-3f model on the IMC dataset.

False Negatives

The ResidualATNet-3f model fails to identify 1492 ATs in the EFilms dataset. In contrast, in the IMC dataset, the ResidualATNet-3f model misses 664 ATs. Examples of missed ATs by the ResidualATNet model i.e. false negatives are provided in Figure 5.18 and Figure 5.19. The captions of the figures also include the similarity values that the

Model	Dataset	Precision	Recall	F1-score	TP	FP	FN
ResidualATNet-3f	EFilms	0.94	0.77	0.85	5009	330	1492
ResidualATNet-3f	**EFilms (corrected)**	**0.95**	**0.77**	**0.85**	**5084**	**255**	**1492**
ResidualATNet-3f	IMC	0.94	0.85	0.90	3805	237	664
ResidualATNet-3f	**IMC (corrected)**	**0.97**	**0.86**	**0.91**	**3926**	**116**	**664**

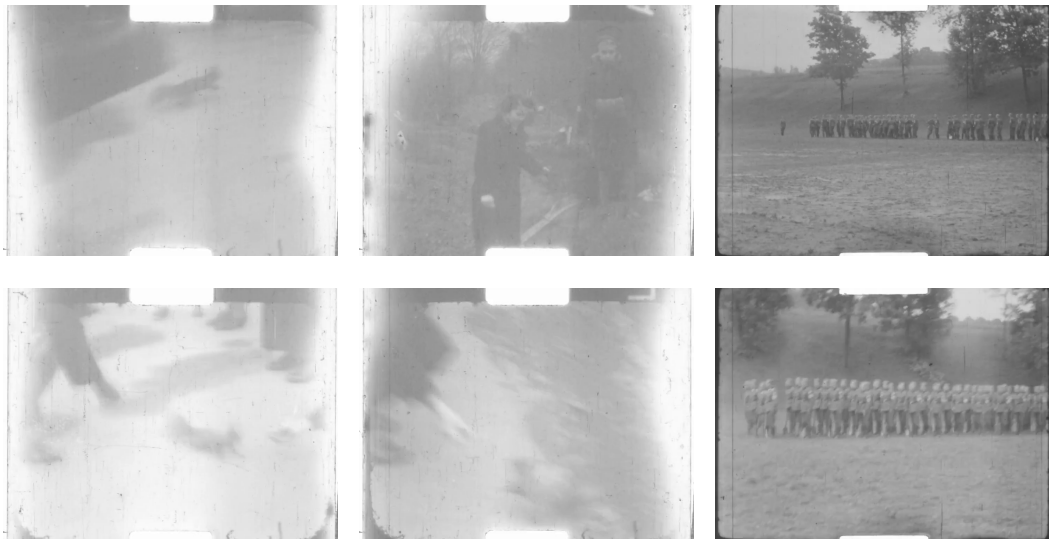
Table 5.12: Overall results achieved by the ResidualATNet-3f model on the EFilms and IMC dataset using the corrected labels.

ResidualATNet model calculates for each pair of false negatives. Since the ResidualATNet model uses an adaptive threshold, the calculated similarity is compared against a threshold value which varies from video to video.

As can be observed in the sample pairs, the historical datasets EFilms and IMC contain very challenging transitions. Figure 5.18a, 5.18b, 5.18c show examples of transitions between two frames which are heavily affected by flicker. Flicker is a very common artefact found in the historical films of EFilms and IMC. It is a consequence of a wrong exposure (overexposure) of the filmstrips to light over the years. In addition to the flicker, the frame pairs in Figure a and b are blurry and show signs of colour fading i.e. low contrast. Artefacts like low contrast, blurriness and flicker interfere with the calculation of the similarity value and contribute to the production of false negatives. The average similarity value of the false-negative samples is over 60%.

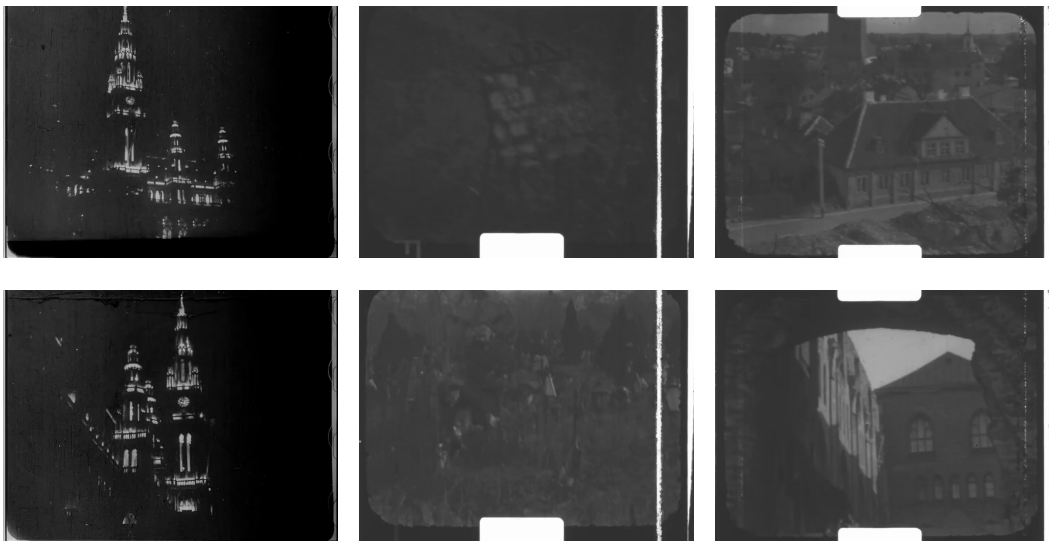
Apart from blur which is also present in sample pairs in Figure 5.19b, 5.19c, another major cause for the not detected transitions in Figure 5.19b, 5.19c is the underexposure of the filmstrips. As can be seen in the examples in Figure 5.19b, 5.19c the frames are blurry with indistinguishable content, vertical scratches and mostly dark. Such dark frames are the result of the underexposure of the filmstrips to light. Vertical scratches and underexposure represent a great challenge for the similarity calculation and consequently pose a serious problem to the AT detection in the historical film material. The similarity value of the frame pairs in Figure 5.19b, 5.19c is over 61% which is why these transitions remain undetected.

Finally, the historical film material of EFilms and IMC stems from the same time period and portray related themes and motives. As a result, the films show very similar content, scenes and settings. Figure 5.18c, 5.19a show examples of similar content and settings. For humans, it is obvious that the frame pairs in Figure 5.18c, 5.19a belong to different shots. However, due to the similarities found in the content, edges and textures in the frames, the ResidualATNet-3f model outputs a similarity value of over 61% for these two cases. These types of similarities negatively affect the detection of ATs in historical videos and add to the total number of false-negative cases in the evaluation of the EFilms and IMC datasets.



(a) Similarity Value: 0.65. (b) Similarity value: 0.66. (c) Similarity value: 0.61.

Figure 5.18: False negative example pairs with their similarity value.



(a) Similarity value: 0.65. (b) Similarity value: 0.73. (c) Similarity value: 0.63.

Figure 5.19: Another set of false negative transition frames with their similarity value.

5.4.4 Comparison to State-of-the-Art Approaches

In this master thesis, six state-of-the-art approaches are implemented and evaluated on the historical film material of EFilms and IMC. The specifics and algorithm details of the approaches selected for the evaluation are provided in Section 2. Out of the six selected methods, two methods are based on conventional SBD techniques. One of the conventional SBD approaches represents an edge-based method which relies on the Edge Change Ratio and follows the algorithm developed by Zabih et al [ZMM95]. The other traditional SBD method is more recent and is based on the multilevel difference of colour histograms. The method is implemented according to the paper by Li et al [LLZ16] and will be referred to as the "Histogram" approach throughout the section.

The rest of the selected state-of-the-art approaches rely on deep learning and utilize CNN architectures for SBD detection. In this work, a basic implementation of DeepSBD [1] is utilized and evaluated with an input size of 16 frames and an overlap of 8 frames. Furthermore, the evaluation of the historical films also takes the implementation of the RidiculouslyFastSBD [2] detector by Gygli [Gyg18] as well as the implementation of the state-of-the-art SBD detector DSM [3] by Tang et al. [TFK+18] into consideration. Finally, in the state-of-the-art evaluation, implementation of CosimNet [4] is included. CosimNet represents a Siamese network [BGL+93] developed for scene change detection by Guo et al. [GFZ+18].

A literature review shows that the state-of-the-art approaches produce an exceptional SBD performance on contemporary film material (see [HES+17] [Gyg18] [TFK+18]). However, the results presented in Table 5.13 and Table 5.14 show that the state-of-the-art approaches face difficulties with the task of SBD when applied to historical film material. This is due to the fact that the deep learning-based approaches are designed, trained and fine-tuned to detect shot transitions in modern high-quality videos. Furthermore, the approaches selected are not implemented with the goal to provide support for grayscale films or cope with the vast number of artefacts found in historical videos. As a result, the approaches struggle on the historical datasets and deliver an overall suboptimal SBD performance.

Table 5.13 summarizes the SBD performance of the state-of-the-art approaches on the EFilms dataset. According to the results presented in Table 5.13, the state-of-the-art approaches can be categorized into three types of performance groups. The first group of approaches includes DSM and CosimNet. The performance of these two approaches is characterized by high precision values and significantly low recall values. On the one hand, high precision values correlate to a low number of false-positive predictions. On the other hand, dramatically low recall values mean that both approaches struggled with identifying the ATs in historical films.

[1] https://github.com/Tangshitao/ClipShots - last accessed: 09.08.2020
[2] https://github.com/abramjos/Scene-boundary-detection - last accessed: 09.08.2020
[3] https://github.com/BitFloyd/deepsbd - last accessed: 09.08.2020
[4] https://github.com/gmayday1997/SceneChangeDet - last accessed: 09.08.2020

Approach	Precision	Recall	F1-score
ECR [ZMM95]	0.58	0.64	0.50
Histogram [LLZ16]	0.63	0.61	0.56
DeepSBD [HES+17]	0.63	0.98	0.76
RidiculouslyFastSBD [Gyg18]	**0.59**	**0.78**	**0.66**
DSM [TFK+18]	0.85	0.22	0.33
CosimNet [GFZ+18]	0.82	0.26	0.36
ResidualATNet-3f	**0.95**	**0.776**	**0.85**

Table 5.13: Results of the state-of-the-art approaches on the EFilms dataset.

Approach	Precision	Recall	F1-score
ECR [ZMM95]	0.54	0.75	0.45
Histogram [LLZ16]	0.65	0.54	0.55
DeepSBD [HES+17]	0.59	0.96	0.72
RidiculouslyFastSBD [Gyg18]	**0.50**	**0.93**	**0.64**
DSM [TFK+18]	0.81	0.32	0.44
CosimNet [GFZ+18]	0.67	0.47	0.51
ResidualATNet-3f	**0.97**	**0.86**	**0.91**

Table 5.14: Results of the state-of-the-art approaches on the IMC dataset.

In the second group, the performance of the approaches has the complete opposite characteristics. For example, DeepSBD achieves a recall value of 98% and a precision value of 63%. This means that DeepSBD identifies almost all transitions, but at the same produces a significant number of false-positive predictions. The detection performance of the approaches is due to the historical film artefacts such as low contrast, shaking, flicker and tears. Finally, the last group involves approaches whose performance is not characterized by a major imbalance in the precision and recall values. Examples of such approaches are ECR and Histogram. ECR and Histogram achieve an F1-score of 50% and 56% respectively. The ECR method is sensitive to small content changes as well as camera and object movements which can cause false detections. As can be observed in Table 5.13, out of the six approaches, on the EFilms dataset, the highest F1-score of 66% is achieved by the approach by the RidiculouslyFastSBD detector.

The results of the state-of-the-art evaluation on the IMC dataset are summarized in Table 5.14. All of the state-of-art approaches produce a similar performance pattern as on the EFilms dataset. When compared to the rest of the state-of-the-art approaches, the RidiculouslyFastSBD detector once again produced most favourable results with an F1-score of 64%. The results of CosimNet and DSM show no significant difference and are characterised with high precision and low recall values. However, it must be

pointed out that CosimNet is not specifically built for SBD but scene change detection. Nonetheless, it produces competitive results when compared to the DSM approach. The RidiculouslyFastSBD detector by Gygli [Gyg18] produces superior results on both datasets due to the utilization of 3D convolutions instead of 2D convolutions. The 3D input of Gygli's model (segments of 100 frames) allows the preservation of contextual spatiotemporal information and thus, enables the analysis of changes over time. The results in Table 5.13 and Table 5.14 show that the state-of-the-art approaches do not support SBD in historical films by default and fine-tuning is required to improve the detection performance.

In comparison to ResidualATNet, the performance of the RidiculouslyFastSBD detector also comes up short on both datasets. As evident in Table 5.13 and Table 5.14, the ResidualATNet-3f model with an adaptive threshold by far outperforms the RidiculouslyFastSBD detector. ResidualATNet-3f provides outstanding performance on historical films and achieves an F1-score of 85% on the EFilms dataset, and 91% on the IMC dataset. The results indicate that in contrast to the state-of-the-art approaches, ResidualATNet-3f successfully deals with the challenges of historical film material. Overall, ResidualATNet-3f provides full support for historical films and enables the SBD in videos and films from the historical domain.

5.4.5 Benchmark datasets Evaluation

The proposed approach is additionally evaluated on contemporary film material from specialized benchmark SBD datasets. The goal of the benchmark dataset evaluation is to assess the efficiency of the proposed approach on modern films and demonstrate its robustness and validity. Note that the proposed approach and the developed CNN models are neither trained nor optimized for contemporary film material. Furthermore, the benchmark dataset evaluation is performed using the same models, weights, hyperparameters as well as threshold values and policies as in the historical film material evaluation.

In the literature, there exist a number of publicly available contemporary film datasets which are used as a benchmark for the assessment of the quality and performance of SBD approaches (see [SOD10] [TFK+18]). The most prominent examples of benchmark SBD datasets include the TRECVID dataset series [SOD10], RAI [BGC15b] and ClipShots [TFK+18]. The proposed SBD method in this work is evaluated on three publicly available benchmark datasets which include RAI [BGC15b], BBC Planet Earth [BGC15a] and Clipshots [TFK+18].

RAI

The RAI [BGC15b] dataset represents a collection of ten randomly selected broadcast videos from the RAI Scuola video archive. The majority of the videos in the RAI dataset consist of talk shows and documentaries. The shot transitions in the RAI database are manually annotated by human experts and includes 721 ATs. The AT detection results

	Fixed threshold			Adaptive threshold		
Video	Precision	Recall	F1-score	Precision	Recall	F1-score
V1 (23553)	0.99	1	0.99	1	1	1
V2 (23557)	1	1	1	1	0.98	0.99
V3 (23558)	0.97	1	0.98	0.99	0.96	0.97
V4 (21867)	0.97	0.90	0.94	0.96	0.91	0.94
V5 (21829)	1	0.91	0.95	0.24	1	0.39
V6 (25008)	1	1	1	0.89	1	0.94
V7 (25009)	0.96	0.97	0.97	1	0.97	0.99
V8 (25010)	0.92	0.99	0.95	0.98	0.90	0.94
V9 (2011)	0.83	0.96	0.89	0.74	0.96	0.83
V10 (25012)	0.94	0.97	0.96	0.83	0.97	0.89
Overall	**0.96**	**0.98**	**0.97**	0.92	0.95	0.93
Mean	**0.96**	**0.97**	**0.96**	0.86	0.97	0.89

Table 5.15: Results achieved by the ResidualATNet-3f model using a fixed threshold and adaptive threshold on the RAI dataset.

achieved by the ResidualATNet-3f architecture on the RAI dataset are summarized in Table 5.15.

Table 5.15 presents the detection performance of the ResidualATNet-3f model per video with a fixed threshold and an adaptive threshold evaluation respectively. The results demonstrate that the ResidualATNet-3f is very robust and can be utilized for AT detection in contemporary film material even though it has not been trained or fine-tuned to support SBD detection in modern films.

As evident in Table 5.15, on the RAI [BGC15b] dataset, the proposed method achieves precision and recall values of at least 90% on the majority of the videos. Furthermore, Table 5.15 shows that the ResidualATNet-3f model produces an outstanding performance and obtains a mean F1-score value of 96% with a fixed threshold evaluation. When evaluated using an adaptive threshold the ResidualATNet-3f model produces a mean F1-score value of 89% (see Table 5.15). This is remarkable, as in the historical film evaluation the fixed threshold evaluation is never superior to the adaptive threshold evaluation. However, note that the method for calculating the adaptive threshold is implemented and fine-tuned to determine the optimal threshold for historical film material. Furthermore, Table 5.15 demonstrates a decline in the performance in videos with a high number of GTs (such as V5). The GTs in the RAI dataset could also affect the calculation of the adaptive threshold and be the reason for the decline in the performance of the ResidualATNet-3f model with an adaptive threshold.

The performance of the state-of-the-art approaches on the RAI [BGC15b] dataset is provided in Table 5.16. The results in Table 5.16 show that the results achieved by the proposed method deliver a competitive SBD performance to the state-of-the-art approaches. The ResidualATNet-3f with a fixed threshold detects almost all ATs and

Approach	Mean F1-Score
DeepSBD [HES+17]	0.94
Histogram [LLZ16]	0.52
RidiculouslyFastSBD [Gyg18]	0.88
DSM [TFK+18]	0.94
Baraldi et al. [BGC15b]	0.84
ResidualATNet-3f (adaptive)	0.89
ResidualATNet-3f (fix)	**0.96**

Table 5.16: Comparison of the state-of-the-art approaches on the RAI dataset.

	Fixed threshold			Adaptive threshold		
Video	Precision	Recall	F1-score	Precision	Recall	F1-score
Bbc_01	0.98	0.92	0.95	0.95	0.95	0.95
Bbc_02	0.99	0.95	0.96	0.99	0.98	0.98
Bbc_03	0.96	0.91	0.94	0.92	0.94	0.93
Bbc_04	0.96	0.97	0.96	0.98	0.97	0.97
Bbc_05	0.97	0.97	0.97	0.89	0.98	0.93
Bbc_06	0.95	0.98	0.97	0.96	0.98	0.97
Bbc_07	0.94	0.91	0.93	0.98	0.81	0.88
Bbc_08	0.39	0.67	0.50	0.38	0.69	0.49
Bbc_09	0.96	0.96	0.96	0.96	0.96	0.96
Bbc_10	0.84	0.86	0.85	0.57	0.95	0.72
Bbc_11	0.97	0.92	0.94	0.95	0.97	0.96
Overall	**0.88**	**0.91**	**0.90**	0.83	0.93	0.88
Mean	**0.90**	**0.91**	**0.90**	0.87	0.93	0.89

Table 5.17: Results achieved by the ResidualATNet-3f model using a fixed threshold and adaptive threshold on the BBC Planet Earth dataset.

produces an outstanding mean F1-score of 96%. From the results presented in Table 5.16 can be inferred that apart from SBD in historical film material the ResidualATNet-3f model can also be employed for automatic SBD in modern videos and contemporary films.

BBC Planet Earth

The BBC Planet Earth [BGC15a] database is a selection of eleven episodes from the BBC educational documentary series Planet Earth. Each video episode is ca. 50 minutes long. The complete BBC Planet Earth database contains 4706 ATs. The performance of the ResidualATNet-3f model is assessed on the BBC Planet Earth dataset with a fixed and an adaptive threshold policy. Table 5.17 summarizes the results which the model achieves per video.

Approach	Precision	Recall	F1-score
DeepSBD [HES+17]	0.73	0.92	0.81
DSM [TFK+18]	**0.78**	**0.93**	**0.85**
ResidualATNet-3f (adaptive)	0.52	0.93	0.58
ResidualATNet-3f (fix)	0.73	0.83	0.74

Table 5.18: Comparison of the results on the ClipShots dataset.

Similarly, to the evaluation on the RAI dataset, on the BBC Planet Earth dataset, the ResidualATNet-3f model in combination with a fixed threshold produces a slightly superior performance to the adaptive threshold evaluation. On the BBC Planet Earth database, the fixed threshold evaluation yields a mean F1-score of 90%. The adaptive threshold evaluation, on the other hand, reaches an average F1-score value of 89%. This evaluation confirms the remarkable AT detection abilities of the ResidualATNet-3f model in a different film domain. Together with the RAI evaluation, the results prove the robustness and film domain independence of the proposed AT detection method.

ClipShots

ClipShots [TFK+18] is the first large-scale benchmark database for SBD. It includes short films collected from YouTube and Weibo which cover 20 different categories ranging from sports, TV shows, competition highlights, animals as well as hand-made videos. The length of each video varies from one to twenty minutes. The evaluation set of ClipShots consists of 500 videos [TFK+18]. From the 500 video evaluation set, 241 videos are used for AT detection validation and the rest are used for GT detection validation. In this evaluation, we assess the AT detection performance of the proposed approach and therefore utilize the 241 AT evaluation videos of ClipShots. The videos contain a total number of 5876 ATs.

The results of the ClipShots evaluation are provided in Table 5.18. As can be observed from the results in Table 5.18, the fixed threshold evaluation of the proposed method produces an F1-score of 74%. Even though the proposed method does not outperform the other two state-of-the-art approaches, it performs considerably well given the fact that the ResidualATNet-3f model is trained and specialized for transition detection in historical material. The ResidualATNet-3f model is neither familiar with the quality and characteristics of short Youtube films nor optimized to identify the transitions in hand-made videos with a mobile device. It is also noted that the DSM approach includes an initial filtering module whereas the DeepSBD approach employs a post-processing filtering technology. In contrast, the proposed approach in this master thesis adopts none of those methods. Lastly, taking the size of the evaluation dataset as well as the high number of transitions into account, ResidualATNet-3f paired with a fixed threshold produces a competitive detection performance on the ClipShots dataset.

IoU	Precision	Recall	F1-score	TP	FN	FP
0.1	0.22	0.78	0.34	35	10	128
0.3	0.2	0.6	0.26	27	18	136
0.4	0.14	0.51	0.22	23	22	140
0.5	0.11	0.4	0.17	18	27	145
0.7	0.1	0.29	0.12	13	32	150
0.8	0.1	0.27	0.12	12	33	151

Table 5.19: The performance of the ResidualGTNet model at different IoU thresholds

Summary

This section presents the experiments and evaluations of the proposed AT detection method on a set of publicly available benchmark datasets. The evaluations demonstrate the robustness and generalization of the proposed AT detection approach. The approach achieves an F1-score of 96% on the RAI dataset and an F1-score of 90% on the BBC Planet Earth dataset. This shows that the proposed approach has outstanding AT detection abilities which are not bound to historical film material. The remarkable performance on the RAI and BBC Planet Earth datasets proves that the ResidualATNet-3f model can be employed for transition detection in films from versatile domains including news and broadcast videos, documentaries and animal series. On the large scale ClipShots dataset, the ResidualATNet-3f model obtains an F1-score value of 74%. The results achieved on the ClipShots dataset are competitive to the state-of-the-art approaches. The evaluations conducted on the benchmark datasets establish the validity and show the impressive performance of the proposed method on contemporary film material.

5.5 GT Detection Experiments

For the evaluation of the GT detection method, the IoU evaluation measure is utilized. Hence, a GT prediction reported by the ResidualGTNet model is considered true positive if the IoU of the reported transition is over the specified threshold ($IoU > t$). Table 5.19 shows the results of the performance of the ResidualGTNet model on the IMC dataset. The GT performance is measured at different IoU thresholds. Figure 5.20 provides a graphical illustration of how F1-score values of the ResidualGTNet model change at different IoU levels. With an IoU of 0.1, the model achieves a precision and recall values of 22% and 78%, respectively. Out of the total 45 transitions, the model detects 35 transitions (TP) in total and fails to detect 10 transitions (FN). Furthermore, the model produces 128 FP predictions which result in an overall F1-score of 34%.

At an IoU level of 0.4, the ResidualGTNet model achieves an F1-score of 22%. In this case, an increase of 12 can be observed in both FP and FN prediction. In contrast, the number of TP predictions is decreased by 12 and the model correctly detects 23 GTs. Consequently, the precision and recall values also experience a drop and equal to 14%

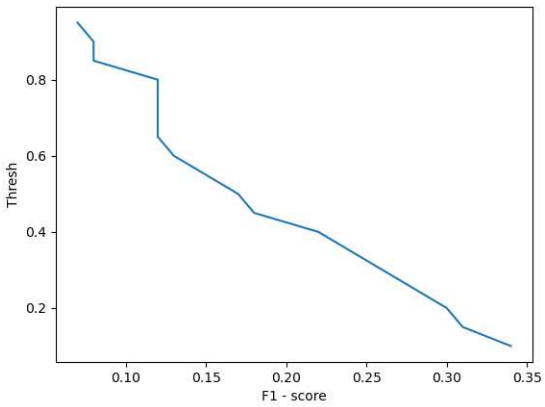


Figure 5.20: The F1-score values over different IoU threshold levels.

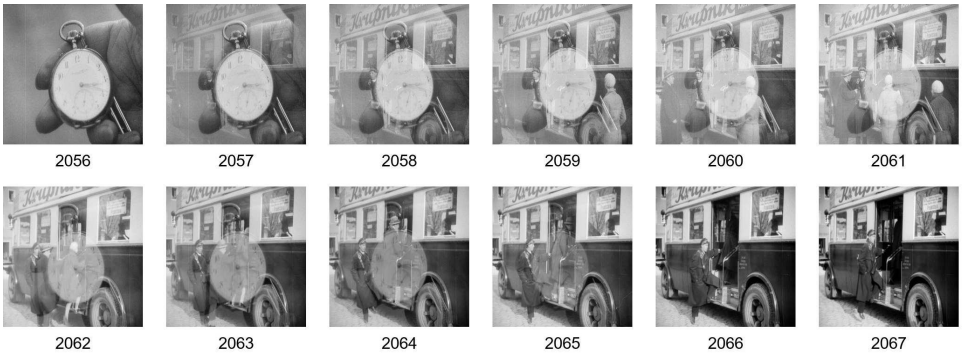
and 51%. On the other hand, when the IoU is configured to 0.5, the model achieves precision and recall values of 11% and 40%. At 0.5 IoU level, the ResidualGTNet model correctly detects only 18 GTs and produces an F1-score of 17%. From the results shown in Table 5.19, it can be deduced that the number of TP and the F1-score value continue to decline at the higher IoU levels.

An analysis of the predictions produced by the ResidualGTNet model shows that the model struggles with three main points. First of all, the ResidualGTModel confuses slow camera and object movements for a GT. This reason is the main contributor to the high number of FP predictions. Secondly, the FP predictions contain cases of flicker, moving shadows and damaged frames. This means that the artefacts found in the historical films have a negative impact on the performance and hinder the GT detection. Thirdly, in the cases where the model locates a GT transition, it also reports a significant number of overhead frames. The results in Table 5.19 show that the GT detection performance decreases in the case of higher IoU thresholds. Consequently, with an IoU threshold of 0.1, the model detects 35 GTs whereas an IoU threshold of 0.7 the model detects only 13 GTs. However, the location of the exact start and end position of a GT is a complex task which requires further analysis and improvement for historical films. Finally, some of the GT found in the historical films are very long and contain multiple effects which makes them extremely difficult to detect.

Figure 5.21 demonstrates examples of the predictions of the ResidualGTNet model. Figure 5.21a and 5.21b both represent examples of correctly identified GT transitions by the ResidualGTNet model. Figure 5.21a has an IoU score of 0.53 which means that it is considered a TP if the threshold is set to 0.5. However, if the threshold is configured to a value higher than 0.5 the prediction in Figure 5.21a is no longer considered correct. On the other hand, the prediction in Figure 5.21b has an IoU score of 0.73 which means the prediction is considered a TP even when the threshold is set to 0.7. An example of a

FN is represented in Figure 5.21c. A reason for this failed detection could be significant similarities between the shots. Finally, Figure 5.21d illustrates the case of slow camera movements which lead to false predictions of the model.

(a) A correctly identified Iris Wipe transition by the ResidualGTNet model with an IoU score of 0.53.



(b) A correctly located Dissolve transition by the ResidualGTNet model with an IoU score of 0.83.

(c) An example of a missed Dissolve transition.

(d) A false positive prediction of the ResidualGTNet model. The model confuses slow camera movements for GTs.

Figure 5.21: Prediction examples of the ResidualGTNet model.

CHAPTER 6

Conclusion and Future Work

In this work, a novel framework for automatic detection of the shot boundaries in historical film material was designed and implemented. The framework is based on deep learning and relies on two separate CNN models for the AT and GT detection. The historical film material utilised in this thesis is related to National socialism and the Holocaust and is divided into two datasets referred to as EFilms and IMC.

For the task of AT detection, two different CNN architectures called ATNet and ResidualATNet were designed and implemented. ATNet and ResidualATNet were trained using three self-generated dataset configurations. The training datasets were created to investigate the effect of the training data colour space (RGB vs. Grayscale) and the data augmentation techniques on the overall SBD performance in historical films. The models were evaluated using two types of feature extraction strategies: single feature extraction and a multi-feature extraction strategy. The multi-feature extraction strategy was employed to examine the influence of multi-feature information on the AT detection process. The contribution of additional information was further examined by extracting information from pre-defined target areas of each video frame. Finally, the transition classification stage was implemented and tested using a fixed and adaptive threshold. The GT detection method, on the other hand, includes a GT candidate selection module and a separate deep CNN model called ResidualGTNet. The selection module was developed to produce a set of transition candidates which are later processed by ResidualGTNet for the actual GT detection. The GT detection is performed using a targeted dissolve detector and a separate fade in/out and wipe detector.

The main objective of this thesis is to address the following research question: *Q1: How do the CNN architectural properties and feature extraction strategies affect the SBD efficiency in historical films?*. From the training experimental studies conducted, it can be concluded that for the task of SBD in historical data, the CNN architecture of ResidualATNet is superior to the CNN architecture of ATNet. Furthermore, the results of the inference experiments showed that the utilisation of a combination of features

referred to as *multi-feature extraction strategy* significantly improves the SBD efficiency. The findings of the training experiments also provided answers to the second research question *Q2: How does the SBD performance in historical films depend on the training data?*. The results of the training experiments indicated that the application of flips as data augmentation technique has no effect on the overall transition detection performance in historical films. In addition, the results demonstrated no significant difference in the performance of the models trained on RGB dataset and the models trained on Grayscale dataset. This finding is attributed to the fact that the majority of the samples in the original HistoricalDataset (RGB) were already in the grayscale colourspace. Finally, the ResidualATNet model trained on the HistoricalDataset and paired with the multi-feature extraction strategy produced an outstanding performance. ResidualATNet achieved 85% F1-score on the EFilms dataset and 91% F1-score on the IMC dataset. In terms of speed, this configuration of ResidualATNet achieved a 28 x real-time speed up and reached an inference speed of 454 FPS on a Tesla P100 - 16GB.

The third research question *Q3: What are the main challenges of historical videos concerning the problem of SBD?* was addressed by the means of a detailed qualitative and quantitative analysis of the historical films and the predictions ResidualATNet model. The analysis revealed a large number of special properties i.e. artefacts of historical films. Filmstrip contraction and tears, shaking, blur, scratches and wrong exposure are just a subset of the artefacts found in the historical film material. The examination of the transition predictions showed that the quality of the films and their artefacts are the main cause for false positive as well as false negative predictions. This kind of artefacts seriously interfere with the transition detection abilities of the model and therefore amplify the complexity of the task of SBD.

To answer the fourth and final research question *Q4: How efficient are the state-of-the-art SBD approaches on historical data?* 2 traditional and 4 state-of-the-art SBD approaches were evaluated on the same test historical film material. In comparison to the considered set of established SBD approaches, the proposed ResidualATNet model produced superior AT detection performance and outperformed both the traditional and state-of-the-art approaches on the historical film material. Out of the six SBD approaches evaluated on the historical films, the most favourable performance was delivered by the deep learning-based RidiculoslyFastSBD by Gygli. RidiculoslyFastSBD by Gygli achieved a 66% F1-score on the EFilms dataset and 64% F1-score on the IMC dataset. The suboptimal performance of the selected SBD approaches was attributed to the artefacts of the historical films which as described above increase the difficulty of the transition detection. Furthermore, the evaluation of the state-of-the-art approaches showed that historical films require special attention and fine-tuning. On the other hand, the high F1-score values achieved by the proposed ResidualATNet model on the historical data confirmed that the model successfully counteracts the artefacts and challenges in historical films which makes it suitable for performing automatic SBD in the historical film domain.

The performance of the best-performing ResidualATNet model is validated on the contemporary film material from three publicly available benchmark datasets RAI, BBC

Planet Earth and ClipShots. ResidualATNet achieved AT detection F1-score of 96% on the RAI dataset and 90% F1-score on the BBC Planet Earth dataset. This performance demonstrated the flexibility and robustness of the proposed approach and showed that its SBD capabilities are not limited to films from the historical domain. On the large-scale ClipShots dataset, the ResidualATNet model achieved an F1-score of 74%. The approach did not outperform the state-of-the-art approaches like DSM and DeepSBD which reached an F1-score of 85% and 81% on the ClipShots dataset. Nevertheless, the results of ResidualATNet are considered competitive taking into account the size of the ClipShots dataset and the fact that ResidualATNet was neither trained nor optimized for modern film material.

The GT detection method was tested and evaluated on the IMC dataset which consists of 45 GTs. The ResidualGTNet model was evaluated using different IoU scores in the range of 0.1 to 0.8. The model achieved 22% precision, 78% recall and 34% overall F1-score at an IoU level of 0.1. Moreover, the ResidualGTNet model detected 35 transitions (TP) out of the 45 GTs and reported 128 false positive detections. With an IoU of 0.5 model achieved an F1-score of 17% and detected 18 GTs. Lastly, at an IoU level of 0.7, the model achieved 12% F1-score and detected only 13 GTs.

The current SBD framework is primarily focused and highly optimized for AT detection and produces outstanding detection results in the historical and contemporary film material. Therefore, future work should aim to optimize and improve the GT detection performance. This can be achieved by developing a new unified GT detection approach. The unified GT detection approach would require the design and implementation of a new CNN model which utilizes 3D convolutions instead of 2D convolutions. This way the model can be directly trained with video segments representing all types of special effects i.e. transitions. As a result, the GT detection method would be generalizable and robust and would be able to detect GTs not only in historical films but also in video material of versatile domains. Another focus point for future work can be the improvement of the inference speed. Even though the current SBD framework runs at 28 x real-time on the historical data, this rate can be further increased by integrating an initial filtering module. The purpose of the initial filtering module would be to select a subset of video frames instead of performing complex computations for each video frame. The adoption of such a pre-processing method will lower the computation costs and increase the overall inference speed of the framework.

List of Figures

List of Tables

List of Algorithms

Bibliography

[ALBK09] Donald A. Adjeroh, M. C. Lee, Nagamani Banda, and Uma Kandaswamy. Adaptive edge-oriented shot boundary detection. *EURASIP J. Image and Video Processing*, 2009, 2009.

[AM14] Evlampios E. Apostolidis and Vasileios Mezaris. Fast shot segmentation combining global and local visual descriptors. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 6583–6587. IEEE, 2014.

[ARS+18] Sadiq H. Abdulhussain, Abd. Rahman Ramli, M. Iqbal Saripan, Basheera M. Mahmmod, Syed Abdul Rahman Al-Haddad, and Wissam A. Jassim. Methods and challenges in shot boundary detection: A review. *Entropy*, 20(4):214, 2018.

[BCS+05] Jesús Bescós, Guillermo Cisneros, José María Martínez Sanchez, José M. Menéndez, and Julián Cabrera. A unified model for techniques on video-shot transition detection. *IEEE Trans. Multimedia*, 7(2):293–307, 2005.

[Ben12] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 437–478. Springer, 2012.

[BGC15a] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. A deep siamese network for scene detection in broadcast videos. In Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan, editors, *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 1199–1202. ACM, 2015.

[BGC15b] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Shot and scene detection via hierarchical clustering for re-using broadcast video. In *Computer Analysis of Images and Patterns - 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015 Proceedings, Part I*, pages 801–811, 2015.

[BGG99] Patrick Bouthemy, Marc Gelgon, and Fabrice Ganansia. A unified approach to shot change detection and camera motion characterization. *IEEE Trans. Circuits Syst. Video Techn.*, 9(7):1030–1044, 1999.

[BGL+93] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a siamese time delay neural network. In Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, editors, *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, pages 737–744. Morgan Kaufmann, 1993.

[BL03] Léon Bottou and Yann LeCun. Large scale online learning. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 217–224. MIT Press, 2003.

[BR96] John S. Boreczky and Lawrence A. Rowe. Comparison of video shot boundary detection techniques. *J. Electronic Imaging*, 5(2):122–128, 1996.

[Can86] John F. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

[CC19] Ingo Zechner (Project Coordinator) and Michael Loebenstein (Deputy Project Coordinator). Ludwig Boltzmann Institute for History and Society and Austrian Film Museum. Project: Visual History of the Holocaust: Rethinking Curation in the Digital Age. `https://www.vhh-project.eu/`, 2019. [Online; last accessed 31.08.2020].

[CFAC03] Matthew Cooper, Jonathan Foote, John Adcock, and Sandeep Casi. Shot boundary detection via similitary analysis. In Alan F. Smeaton, Wessel Kraaij, and Paul Over, editors, *2003 TREC Video Retrieval Evaluation, TRECVID 2003, Gaithersburg, MD, USA, November 17-18, 2003*. National Institute of Standards and Technology (NIST), 2003.

[Cha91] W. Chauvenet. *A Manual of Spherical and Practical Astronomy: Embracing the General Problems of Spherical Astronomy, the Special Applications to Nautical Astronomy, and the Theory and Use of Fixed and Portable Astronomical Instruments, with an Appendix on the Method of Least Squares. ...* A Manual of Spherical and Practical Astronomy: Embracing the General Problems of Spherical Astronomy, the Special Applications to Nautical Astronomy, and the Theory and Use of Fixed and Portable Astronomical Instruments, with an Appendix on the Method of Least Squares. J.B. Lippincott, 1891.

[CLL+15] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *CoRR*, abs/1512.01274, 2015.

[CS02] Barbara Clark and Susan J. Spohr. *Guide to postproduction for TV and film : managing the process.* Focal Press, Amsterdam ; Boston, 2nd ed.. edition, 2002.

[EGW+10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

[Fos18] Giovanna Fossati. *From Grain to Pixel - The Archival Life of Film in Transition.* Amsterdam University Press, 2018.

[GFZ+18] Enqiang Guo, Xinsha Fu, Jiawei Zhu, Min Deng, Yu Liu, Qing Zhu, and Haifeng Li. Learning to measure change: Fully convolutional siamese metric networks for scene change detection. *CoRR*, abs/1810.09111, 2018.

[GGD12] Lakshmi Priya G G and S. Domnic. Edge strength extraction using orthogonal vectors for shot boundary detection. *Procedia Technology*, 6:247–254, 12 2012.

[GKS00] Ullas Gargi, Rangachar Kasturi, and Susan H. Strayer. Performance characterization of video-shot-change detection methods. *IEEE Trans. Circuits Syst. Video Techn.*, 10(1):1–13, 2000.

[Gyg18] Michael Gygli. Ridiculously fast shot boundary detection with fully convolutional neural networks. In *2018 International Conference on Content-Based Multimedia Indexing, CBMI 2018, La Rochelle, France, September 4-6, 2018*, pages 1–4, 2018.

[Han02] Alan Hanjalic. Shot-boundary detection: unraveled and resolved? *IEEE Trans. Circuits Syst. Video Techn.*, 12(2):90–105, 2002.

[HCL06] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society, 2006.

[HES+17] Ahmed Hassanien, Mohamed A. Elgharib, Ahmed Selim, Mohamed Hefeeda, and Wojciech Matusik. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks. *CoRR*, abs/1705.03281, 2017.

[HK19a] Daniel Helm and Martin Kampel. Shot boundary detection for automatic video analysis of historical films. In Marco Cristani, Andrea Prati, Oswald Lanz, Stefano Messelodi, and Nicu Sebe, editors, *New Trends in Image Analysis and Processing - ICIAP 2019 - ICIAP International Workshops, BioFor, PatReCH, e-BADLE, DeepRetail, and Industrial Session, Trento, Italy, September 9-10, 2019, Revised Selected Papers*, volume 11808 of *Lecture Notes in Computer Science*, pages 137–147. Springer, 2019.

[HK19b] Daniel Helm and Martin Kampel. Video shot analysis for digital curation and preservation of historical films. In Selma Rizvic and Karina Rodriguez-Echavarria, editors, *GCH 2019 - Eurographics Workshop on Graphics and Cultural Heritage, GCH 2019, Sarajevo, Bosnia and Herzegovina, November 6-9, 2019*, pages 25–28. Eurographics Association, 2019.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[IZ16] Michael Loebenstein Ingo Zechner. Ludwig Boltzmann Institute for History and Society and Austrian Film Museum: I-Media-Cities. `https://imediacities.hpc.cineca.it/app/catalog`, 2016. [Online; last accessed 31.08.2020].

[KGU10a] Onur Küçüktunç, Ugur Güdükbay, and Özgür Ulusoy. Fuzzy color histogram-based video segmentation. *Comput. Vis. Image Underst.*, 114(1):125–134, 2010.

[KGU10b] Onur Küçüktunç, Ugur Güdükbay, and Özgür Ulusoy. Fuzzy color histogram-based video segmentation. *Computer Vision and Image Understanding*, 114(1):125–134, 2010.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[LBD+89] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In David S. Touretzky, editor, *Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]*, pages 396–404. Morgan Kaufmann, 1989.

[Lip19] Stan Lipovetsky. Digital humanities and film studies: Visualising dziga vertov's work. *Technometrics*, 61(4):562–563, 2019.

[LLZ16] Z. Li, X. Liu, and S. Zhang. Shot boundary detection based on multilevel difference of colour histograms. In *2016 First International Conference on Multimedia and Image Processing (ICMIP)*, pages 15–22, June 2016.

[LS13] Zhe-Ming Lu and Yong Shi. Fast video shot boundary detection based on SVD and pattern matching. *IEEE Trans. Image Process.*, 22(12):5136–5145, 2013.

[LZ01] Rainer Lienhart and André Zaccarin. A system for reliable dissolve detection in videos. In *Proceedings of the 2001 International Conference on Image Processing, ICIP 2001, Thessaloniki, Greece, October 7-10, 2001*, pages 406–409. IEEE, 2001.

[MF03] Jordi Mas and Gabriel Fernandez. Video shot boundary detection based on color histogram. In Alan F. Smeaton, Wessel Kraaij, and Paul Over, editors, *2003 TREC Video Retrieval Evaluation, TRECVID 2003, Gaithersburg, MD, USA, November 17-18, 2003*. National Institute of Standards and Technology (NIST), 2003.

[MMK+19] Markus Mühling, Manja Meister, Nikolaus Korfhage, Jörg Wehling, Angelika Hörth, Ralph Ewerth, and Bernd Freisleben. Content-based video retrieval in historical collections of the german broadcasting archive. *Int. J. on Digital Libraries*, 20(2):167–183, 2019.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119, 2013.

[OR06] N. Ohta and A. Robertson. *Colorimetry: Fundamentals and Applications*. The Wiley-IS&T Series in Imaging Science and Technology. Wiley, 2006.

[PS16] K. K. Pal and K. S. Sudeep. Preprocessing for image classification by convolutional neural networks. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 1778–1781, 2016.

[QWG03] Laiyun Qing, Weiqiang Wang, and Wen Gao. Illumination invariant shot boundary detection. In Jiming Liu, Yiu-ming Cheung, and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL 2003, Hong Kong, China, March 21-23, 2003, Revised Papers*, volume 2690 of *Lecture Notes in Computer Science*, pages 1097–1101. Springer, 2003.

[SL20] Tomás Soucek and Jakub Lokoc. Transnet V2: an effective deep network architecture for fast shot transition detection. *CoRR*, abs/2008.04838, 2020.

[SOD10] Alan F. Smeaton, Paul Over, and Aiden R. Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.

[SZ15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[SZL15] Ling Shao, Fan Zhu, and Xuelong Li. Transfer learning for visual categorization: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 26(5):1019–1034, 2015.

[SZMB11] Markus Seidl, Matthias Zeppelzauer, Dalibor Mitrovic, and Christian Breiteneder. Gradual transition detection in historic film material - a systematic study. *JOCCH*, 4(3):10:1–10:18, 2011.

[TBF+15] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4489–4497, 2015.

[TFK+18] Shitao Tang, Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang. Fast video shot transition localization with deep structured models. In *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part I*, pages 577–592, 2018.

[UGE06] Oguzhan Urhan, M. Kemal Güllü, and Sarp Ertürk. Modified phase-correlation based robust hard-cut detection with application to archive film. *IEEE Trans. Circuits Syst. Video Techn.*, 16(6):753–770, 2006.

[VW02] Jeroen Vendrig and Marcel Worring. Systematic evaluation of logical story unit segmentation. *IEEE Trans. Multimedia*, 4(4):492–499, 2002.

[WZJ+19] Lifang Wu, Shuai Zhang, Meng Jian, Zhe Lu, and Dong Wang. Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks. *IEEE Access*, 7:77268–77276, 2019.

[XSX16] Jingwei Xu, Li Song, and Rong Xie. Shot boundary detection using convolutional neural networks. In *2016 Visual Communications and Image Processing, VCIP 2016, Chengdu, China, November 27-30, 2016*, pages 1–4, 2016.

[YWX+07] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang. A formal study of shot boundary detection. *IEEE Trans. Circuits Syst. Video Techn.*, 17(2):168–186, 2007.

[Zec15] Ingo Zechner. Ludwig Boltzmann Institute for History and Society: Ephemeral Films Project National Socialism in Austria. `http://efilms.ushmm.org/`, 2015. [Online; last accessed 31.08.2020].

[ZKS93] HongJiang Zhang, Atreyi Kankanhalli, and Stephen W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Syst.*, 1(1):10–28, 1993.

[ZMB08] Matthias Zeppelzauer, Dalibor Mitrovic, and Christian Breiteneder. Analysis of historical artistic documentaries. In *Ninth International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2008, Klagenfurt, Austria, May 7-9, 2008*, pages 201–206. IEEE Computer Society, 2008.

[ZMB12] Matthias Zeppelzauer, Dalibor Mitrovic, and Christian Breiteneder. Archive film material - a novel challenge for automated film analysis. *The Frames Cinema Journal*, 1(1), 2012.

[ZMM95] Ramin Zabih, Justin Miller, and Kevin Mai. A feature-based algorithm for detecting and classifying scene breaks. In Polle Zellweger, editor, *Proceedings of the Third ACM International Conference on Multimedia '95, San Francisco, CA, USA, November 5-9, 1995*, pages 189–200. ACM Press, 1995.

[ZMZB11] Maia Zaharieva, Dalibor Mitrovic, Matthias Zeppelzauer, and Christian Breiteneder. Film analysis of archived documentaries. *IEEE Multim.*, 18(2):38–47, 2011.

[ZZMB10] Matthias Zeppelzauer, Maia Zaharieva, Dalibor Mitrovic, and Christian Breiteneder. A novel trajectory clustering approach for motion segmentation. In Susanne Boll, Qi Tian, Lei Zhang, Zili Zhang, and Yi-Ping Phoebe Chen, editors, *Advances in Multimedia Modeling, 16th International Multimedia Modeling Conference, MMM 2010, Chongqing, China, January 6-8, 2010. Proceedings*, volume 5916 of *Lecture Notes in Computer Science*, pages 433–443. Springer, 2010.