


Video Shot Analysis for Digital Curation and Preservation of Historical Films

D. Helm¹  and M. Kampel¹ 

¹TU Wien, Institute for Visual Computing and Human-Centered Technology, Austria

Abstract

In automatic video analysis and film preservation, Shot Boundary Detection (SBD) and Shot Type Classification (STC) are fundamental pre-processing steps. While previous research focuses on detecting and classifying shots in different video genres such as sports movies, documentaries or news clips only few studies investigate on SBD and STC in historical footage. In order to promote research on automatic video analysis the project Visual History of the Holocaust (VHH) has been started in January 2019. The main aim of this paper is to present first results on the fundamental topics SBD and STC in the context of the project VHH. Therefore, a deep learning-based SBD approach is implemented to detect Abrupt Transitions (ATs). Furthermore, a CNN-based algorithm is analyzed and optimized in order to classify shots into the four categories: Extreme-Long-Shot (ELS), Long-Shot (LS), Medium-Shot (MS) and Close-Up (CU). Finally, both algorithms are evaluated on a self-generated historical dataset related to the National Socialism and the Holocaust. The outcome of this paper demonstrates a first quantitative evaluation of the SBD approach and displays a $F_{1,Score}$ of 0.866 without the need of any re-training or optimization. Moreover, the proposed STC algorithm reaches an accuracy of 0.71 on classifying shots. This paper contributes a significant base for future research on automatic shot analysis related to the project VHH.

CCS Concepts

• **Information systems** → Video search; • **Computing methodologies** → Supervised learning by classification; Transfer learning;

1. Introduction

For decades, main reference points of Europe's political and cultural heritage are Nazi concentration camps and the Holocaust [MrsO19]. Furthermore, the visual representation of the Holocaust plays a central role for film archivists, artists, historians, educators, and curators. Archiving and working with a huge amount of different types of information such as film records, text-based documents and oral histories of contemporary witnesses is very time-consuming and exhausting. One reason is that the data are located in different archives whereas each of them has various digitization and archival standards. A further point is the considerable amount of data, which are unknown or unseen by any expert up to now. Finally, the content of film records related to the discovery of Nazi concentration camps and other atrocity sites is a huge strain for each individual expert [MrsO19]. The H2020 project *Visual History of the Holocaust* has been funded in order to counteract these challenges by developing international digitization standards for historical film preservation of Holocaust related footage as well as automatic video analysis tools. The explored methods accomplish new efficient and innovative possibilities for memorials, educational institutions and museums to create new visual representations of the

Holocaust as well as to interact with Holocaust-related objects or information [MrsO19]. The fundamental base for automatic video analysis tools is Shot Boundary Detection (SBD) and Shot Type Classification (STC) [ZMB12]. SBD as well as STC have been an active research field for decades [SZMB11] [CBL13]. For SBD, studies have focused on traditional computer vision approaches [ZXS16] [ALBK09] [PMT03] in the last decades whereas recent research focuses on Deep Learning techniques [JLR17] [TFK*18] [HK19]. Furthermore, studies on STC focus on classifying shot types such as Long-Shot (LS), Medium-Shot (MS) and Close-Up (CU) by detecting a persons' face and use the ratios between the face size in combination with the frame size to classify one frame in one of these categories [VTNP12]. However, recent research also investigates on Deep Learning-based approaches [MJI*19]. While the detection and classification of shots in video genres such as sports movies or news clips have been the main scope of past research [MJI*19], fewer studies have investigated on historical footage which shows specific challenges such as damaged film reels, scratches, splices or the exposure and occurs during the digitization process of analog films [ZMB12] [SZMB11]. A shot is defined as a number of consecutive frames related to the same content and is triggered by starting as well as ending a recording with

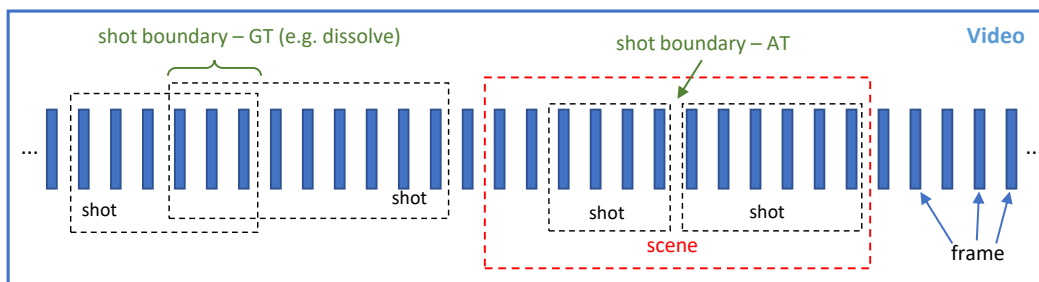


Figure 1: Basic components of a semi-professional or professional video (inspired by [SZMB11]).

one camera [ZXS16]. A further structural level in film-making is the so-called scene. A scene is built by connecting one to several recorded shots with specific boundaries. A professionally or semi-professionally created movie or video clip consists of a varying number of combined scenes which together form the final movie [SZMB11]. Figure 1 visualizes the basic components of a film or video clip [ZMB12]. There are two main types of shot boundaries which are used to connect single shots: Abrupt Transition (AT) and Gradual Transition (GT). ATs are defined by a significant semantic change between two consecutive frames. That means that the next shot appears directly after the last frame of the previous shot. A GT is described as a smoother change of two adjacent shots and includes several frames of both shots. This is achieved by transition techniques such as fades, wipes or dissolves [SZMB11]. Each shot is recorded with specific camera settings such as the distance between the object and the camera. In professional film making there exist eight standard types [MJJ*19]: Extreme-Long-Shot (ELS), Long-Shot (LS), Medium-Full-Shot (MFS), American Shot (AS), Medium-Shot (MS), Medium-Close-Shot (MCS), Close-Up (CU) and Extreme-Close-Up (ECU). The distance to an object, mostly a person, is used to give a recorded shot additional information. This paper demonstrates SBD and STC algorithms in order to detect and classify shots in historical videos related to the project VHH. Finally, it demonstrates that the evaluated SBD and STC mechanisms form a significant and fundamental base for future research.

The structure of this paper is organized as follows: Section 2 introduces the project VHH. An overview of the state-of-the-art is demonstrated in Section 3. In Section 4 the methodology and first results of SBD and STC, are presented. Finally, the paper closes with the conclusion in Section 5.

2. Project: Visual History of the Holocaust

The Horizon 2020 (H2020) project Visual History of the Holocaust (VHH) is funded by the European Union, has started in January 2019 and takes four years.

2.1. Main Objectives

The main objectives are to develop and define international digitization standards for libraries and archives in order to preserve historical films recorded during the discovery of Nazi concentration camps. Due to the enormous number of available information

as films, audio recordings and text-based documents, a web application, called VHH Media Management and Search Infrastructure (VHH-MMSI) is developed in order to provide historians, educational institutions, libraries and archives new innovative ways to work with digitized information related to the Holocaust (see illustration in Fig. 2). The core of application is represented by the automatic video analysis tools which are split in five topics:

- Shot Boundary Detection - SBD
- Shot Type Classification - STC
- Camera Movement Detection - CMD
- Object Detection - OD
- Relation Detection - RD

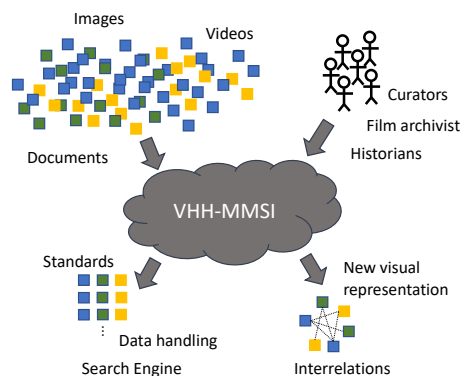


Figure 2: Schematic illustration of the VHH-MMSI system.

2.2. Partners and Responsibilities

The joined efforts of experts in different fields such as technology, film archiving and history are required in order to achieve the presented goals. The project consortium is formed by 14 partners. The project is coordinated by the Ludwig Boltzmann Institute (LBI - Ingo Zechner) and the Austrian Film Museum (OFM - Michael Löwenstein). Figure 3 demonstrates a rough schematic visualization of the partner institutions.

3. State-of-the-art

Shot Boundary Detection: SBD tends to provide more promising results using deep learning approaches [Gyg18] [TFK*18]

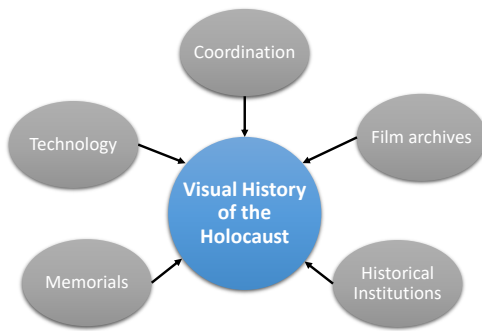


Figure 3: Schematic illustration of the partner core fields of the project VHH.

[JLR17] instead of traditional computer vision algorithms such as histogram-based [ZXS16]. Gygli [Gyg18] presents an algorithm which focuses on detecting shot transitions such as ATs and GTs by using a 3D Convolutional Neural Network. The automatically generated dataset consists of transition snippets as well as of non-transition snippets and can be used to train the proposed 3D-CNN end-to-end. The algorithm is evaluated by using the data set RAI [BGC15]. A further CNN-based SBD technique has been presented by Jingwei et al. [JLR17]. The extracted feature vectors of the pre-trained CNN model for each frame are evaluated by calculating the cosine similarity metric. As a pre-processing step, they have built a candidate segment selection based on an adaptive threshold mechanism. Hassanien et al. [HES*17] have published a shot boundary detection mechanism also based on a 3D Convolutional Neural Network (3D-CNN), called DeepSBD, to detect abrupt as well as gradual transitions in videos. The approach published by Tang et al. [TFK*18] is inspired by [HES*17] and contributes a further fast and accurate SBD method. In a first step the whole video is filtered in order to select candidate segments including shot transitions. The ATs are detected by using a 2D ConvNet which is trained on a similarity function between two images and is able to locate the position of ATs. In order to locate GTs a 3D CNN network based on the idea of DeepSBD [HES*17] is used.

Shot Type Classification: A second literature review shows that Deep Learning techniques also dominate the classification of shot types such as LS, MS, CU. However, also most studies investigate on sports movies, parliamentary debates or modern Hollywood movies whereas only few studies focus on historical films. Halin et. al [HRR09] present an approach to classify the shot types ELS and CU. The algorithm is based on segmentation and splits a given frame into non-playfield and playfield regions. In a final step the object sizes in the playfield regions are analyzed and used to classify into ELS or CU. Minhas et al. [MJI*19] have presented a further approach for STC by using an adapted and retrained AlexNet CNN architecture to classify CU, MS, LS and crowd views of a given shot. Benini et al. [BSA*16] have published an approach to classify shots into three types, MS, LS, CU. The algorithm is based on extracting features of five different domains such as color intensity, motion activity, geometry, image content as well as image spectral components. These features are applied to the classifiers

SVM and Decision Tree in order to assign the given shot in one out of three categories: LS, MS, CU. Vretos et al. [VTNP12] presents a shot classification method based on the centric actors' face. The ratio between height and width of the detected face bounding box and the frame is used to assign the frame to one class out the seven categories: ECU, CU, MCU, MS, MLS, LS and ELS. Therefore, an SVM classifier is applied to the extracted features. Savardi et al. [SSMB18] present a further approach to classify movie frames into one out of the three shot scale categories: LS, MS, CU. Therefore, they have explored the architectures, VGG16, GoogleNet and AlexNet and generated a dataset which includes 400000 frames of 120 different movies.

4. Methodology & Results

A SBD method inspired by [Gyg18] [HES*17] [TFK*18] and a STC based on a pre-trained CNN network architecture [SSMB18] are implemented and evaluated. The evaluation of both approaches is done by using a self-generated historical film dataset [HK19].

4.1. Dataset - EFilms_DB

To evaluate the proposed methodologies a self-generated dataset is created which includes historical films related to the Holocaust and the National-socialism and the films are published during the project Ephemeral Films [ZfGuG15]. The self-generated dataset used in this paper consists of 66 videos with varying numbers of frames. Moreover, the shot boundaries in these videos are annotated by experts of the VHH project consortium. Figure 4 demonstrates few example frames of the dataset [HK19].



Figure 4: Example frames of randomly selected videos of the EFilms_DB.

A first evaluation of the generated dataset and the corresponding annotations shows that there is a significant imbalance between the available number of ATs (7145) and GTs (69) in the videos. Furthermore, also the annotations related to the shot types displays a significant imbalance: ELS (286), LS (3529), MFS (408), AS (135), MS (2180), MCS (504), CU (327) and ECU (5). Therefore, the focus in a first evaluation is on detecting ATs and classifying ELS, LS, MS, and CU.

4.2. SBD

In an earlier investigation [HK19] we have published a new SBD algorithm inspired by [Gyg18] [HES*17] [TFK*18] in order to detect ATs in the EFilms_DB. The method used, shows promising results on detecting ATs by using the combination of state-of-the-art solutions without the need to fine-tune and optimize them. Table 1 demonstrates the results of our published algorithm tested on the EFilms_DB.

Table 1: Comparison of Precision, Recall and $F_{1,Score}$ of Shot Boundary Detection mechanisms.

Method	Precision	Recall	$F_{1,Score}$	Evaluated on
Hassanien et. al [HES*17]	0.944	0.818	0.877	EFilms_DB
Li et al. [ZXS16]	0.571	0.540	0.504	EFilms_DB
own [HK19]	0.895	0.898	0.897	Clipsshots_DB
own [HK19]	0.891	0.841	0.866	EFilms_DB

4.3. STC

A first evaluation investigates on the CNN-based algorithm published by [SSMB18] as base for classifying the shot types: ELS, LS, MS and CU. The authors have published a pre-trained CNN model on a self-generated dataset including the shot types LS, MS and CU for about 400000 frames. Table 2 compares the results on the original model from [SSMB18] trained on modern Hollywood movies with a fine-tuned model on the EFilms_DB in order to classify the four categories: ELS, LS, MS and CU. The fine-tuning process is done by splitting the EFilms_DB into training (60%), validation (20%) and testset (20%).

Table 2: Comparison of Precision, Recall, $F_{1,Score}$ and Accuracy of the evaluated experiments. All experiments are evaluated on the testset of the EFilms_DB.

Method	Precision	Recall	$F_{1,Score}$	Accuracy
Output: LS, MS, CU (Original)	0.571	0.619	0.591	0.619
Fine-tuned (all layers)	0.722	0.705	0.712	0.705

5. Conclusion

This paper demonstrates first results of a SBD algorithm in order to detect ATs in historical films related to the H2020 project VHH. Moreover, it shows a STC mechanism for classifying shots into the categories: ELS, LS, MS and CU. The proposed SBD algorithm demonstrates a $F_{1,Score}$ of 0.866 on detecting ATs without the need of any optimization or re-training. Furthermore, this paper points out that fine-tuning the pre-trained STC algorithm with the self-generated weakly-labeled historical dataset displays an increase of the test accuracy of about 0.08 compared to the original model. Future research focuses on reducing false detections of ATs triggered by scratches, damaged film reels or different exposures. Moreover, a further investigation is on fine-tuning and optimizing the CNN-based models for both: STC and SBD. However, the current result stage forms a significant and fundamental base for future research.

6. Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement No. 822670.

References

[ALBK09] ADJEROH D., LEE M. C., BANDA N., KANDASWAMY U.: Adaptive edge-oriented shot boundary detection. *EURASIP Journal on Image and Video Processing* 2009, 1 (2009), 859371. 1

[BGC15] BARALDI L., GRANA C., CUCCHIARA R.: Shot and scene detection via hierarchical clustering for re-using broadcast video. In *Computer Analysis of Images and Patterns* (Cham, 2015), Azzopardi G., Petkov N., (Eds.), Springer International Publishing, pp. 801–811. 3

[BSA*16] BENINI S., SVANERA M., ADAMI N., LEONARDI R., KOVÁCS A. B.: Shot scale distribution in art films. *Multimedia Tools and Applications* 75, 23 (2016), 16499–16527. 3

[CBL13] CANINI L., BENINI S., LEONARDI R.: Classifying cinematographic shot types. *Multimedia Tools and Applications* 62, 1 (2013), 51–73. 1

[Gyg18] GYGLI M.: Ridiculously fast shot boundary detection with fully convolutional neural networks. *Biochimica et biophysica acta* 89, 1 (2018), 95–108. 2, 3

[HES*17] HASSANIEN A., ELGHARIB M. A., SELIM A., HEFEEDA M., MATUSIK W.: Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks. *CoRR abs/1705.03281* (2017). 3, 4

[HK19] HELM D., KAMPEL M.: Shot boundary detection for automatic video analysis of historical films. In *New Trends in Image Analysis and Processing – ICIAP 2019* (Cham, 2019), Cristani M., Prati A., Lanz O., Messelodi S., Sebe N., (Eds.), Springer International Publishing, pp. 137–147. 1, 3, 4

[HRR09] HALIN A. A., RAJESWARI M., RAMACHANDRAM D.: Shot view classification for playfield-based sports video. *ICSIPA09 - 2009 IEEE International Conference on Signal and Image Processing Applications, Conference Proceedings*, 501 (2009), 410–414. 3

[JLR17] JINGWEI X., LI S., RONG X.: Shot boundary detection using convolutional neural networks. *VCIP 2016 - 30th Anniversary of Visual Communication and Image Processing* (2017), 1–4. 1, 3

[MJJ*19] MINHAS R. A., JAVED A., IRTAZA A., MAHMOOD M. T., JOO Y. B.: Shot Classification of Field Sports Videos Using AlexNet Convolutional Neural Network. *Applied Sciences* 9, 3 (2019), 483. 1, 2, 3

[MrsO19] MOSER A., RTD SERVICES OG: Project: Visual history of the holocaust, 2019. URL: <https://www.vhh-project.eu/en/summary/>. 1

[PMT03] PORTER S., MIRMEHDI M., THOMAS B.: Temporal video segmentation and classification of edit effects. *Image and Vision Computing* 21, 13-14 (2003), 1097–1106. 1

[SSMB18] SAVARDI M., SIGNORONI A., MIGLIORATI P., BENINI S.: Shot Scale Analysis in Movies by Convolutional Neural Networks. *Proceedings - International Conference on Image Processing, ICIP* (2018), 2620–2624. 3, 4

[SZMB11] SEIDL M., ZEPPELZAUER M., MITROVIĆ D., BREITENEDER C.: Gradual transition detection in historic film material - a systematic study. *J. Comput. Cult. Herit.* 4, 3 (2011), 10:1–10:18. 1, 2

[TFK*18] TANG S., FENG L., KUANG Z., CHEN Y., ZHANG W.: Fast video shot transition localization with deep structured models. *CoRR abs/1808.04234* (2018). 1, 2, 3

[VTNP12] VRETOS N., TSINGALIS I., NIKOLAIDIS N., PITAS I.: Svm-based shot type classification of movie content. 1, 3

[ZfGuG15] ZECHNER I., FÜR GESCHICHTE UND GESELLSCHAFT L. B. I.: Project: Ephemeral films project national socialism in austria, 2015. URL: <http://efilms.ushmm.org/>. 3

[ZMB12] ZEPPELZAUER M., MITROVIC D., BREITENEDER C.: Archive film material - a novel challenge for automated film analysis. *The Frames Cinema Journal* 1, 1 (2012). 1, 2

[ZXS16] ZONGJIE L., XIABI L., SHUWEN Z.: Shot Boundary Detection based on Multilevel Difference of Colour Histograms. *Proceedings - 2016 1st International Conference on Multimedia and Image Processing, ICMIP 2016* (2016), 15–22. 1, 2, 3, 4