HistShot: A Shot Type Dataset based on Historical Documentation during WWII

Daniel Helm[a], Florian Kleber[b] and Martin Kampel[c]
Computer Vision Lab, Institute of Visual Computing and Human-Centered Technology, TU Wien, Favoritenstraße 9/193-1, Vienna, Austria

Keywords: Historical Film Preservation, Film Archives, Deep Learning, Automated Film Analysis, Film Shot Dataset.

Abstract: Automated shot type classification plays a significant role in film preservation and indexing of film datasets. In this paper a historical shot type dataset (HistShot) is presented, where the frames have been extracted from original historical documentary films. A center frame of each shot has been chosen for the dataset and is annotated according to the following shot types: Close-Up (CU), Medium-Shot (MS), Long-Shot (LS), Extreme-Long-Shot (ELS), Intertitle (I), and Not Available/None (NA). The validity to choose the center frame is shown by a user study. Additionally, standard CNN-based methods (ResNet50, VGG16) have been applied to provide a baseline for the HistShot dataset.

1 INTRODUCTION

Professional produced films such as modern Hollywood productions as well as films from the 1950s or 1970s are not created by just recording one specific situation (Fossati, 2018; Flückiger et al., 2018). In fact, they consist of a complex film hierarchy (Fossati and van den Oever, 2016) and are produced as well as published by following an editing and recording process. Thus, a professional produced film consists of the following hierarchy: *film*, *scene*, *shot* and *frame*. The smallest unit in a film is represented by the **frame**. Frames are recorded with visual recording mediums such as a handheld camera or professional film recording equipment and contain a captured real-world scene. The next level in the film hierarchy is known as the **shot**. One shot is the basic unit in professionally produced movies and consists of a number of consecutive recorded frames. This means that one shot is determined by triggering the start point and endpoint of a recording with specific camera settings. Multiple recorded shots related to the same situation are edited by cutting some frames at the beginning or end. In the next step, they are concatenated together to form a so-called **scene**. For example, shots corresponding to the same scene show the same subject in one specific situation. However, shots are recorded using different settings such as the camera position or the distance between the camera, and the subject of interest (Luca et al., 2013; Fossati, 2018). Finally, a **film** includes multiple concatenated scenes. An overview of the components of a professional film is visualized in Figure 1. Commonly, movies consist of several shots and scenes.

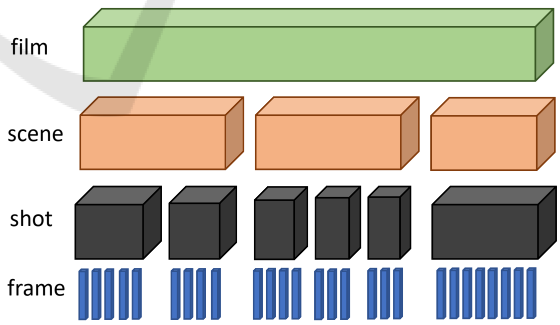

Figure 1: The core components and the hierarchy of a professional produced film are visualized.

Scenes can have cinematographic settings used to characterize individual shots. Two fundamental ones are related to basic camera settings and operations: *shot type* and *camera movement*. The focus in this paper is on the **shot type** or also known as the **shot size**. This characteristic is a kind of representation of the distance between the subject of interest and the camera lens (Luca et al., 2013; Benini et al., 2016).

[a] https://orcid.org/0000-0002-2195-7587
[b] https://orcid.org/0000-0001-8351-5066
[c] https://orcid.org/0000-0002-5217-2854

Helm, D., Kleber, F. and Kampel, M.
HistShot: A Shot Type Dataset based on Historical Documentation during WWII.
DOI: 10.5220/0010872500003122
In *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2022)*, pages 636-643
ISBN: 978-989-758-549-4; ISSN: 2184-4313

One common definition of subcategories of shot types is as follows: Extreme Long Shot (ELS), Long Shot (LS), Medium Long Shot (MLS), Medium Shot (MS), American Shot (AS), Medium Close Shot (MCS), Close Up (CU) and Extreme Close Up (ECU) (see Figure 2 some examples)(Cherif et al., 2007; Zechner and Loebenstein, 2016). Those categories are used to give a recorded situation in a film a specific characteristic. For example, a CU can point out strong emotions of a person related to a specified situation, while an ELS is used to let the viewer dive into the depth of a scene. A(n) (automated) shot type classification allows to analyze films and documentaries. This can be done by e.g., film indexing or content understanding based on the shot type.

This paper proposes a new dataset containing frames extracted from original digitized historical documentaries related to the time of the liberation phase of Nazi-Concentration-Camps during the Second World War (1943-1945). The Historical Film Shot Dataset (HistShot) includes 1885 samples with corresponding frame-based shot type annotations. This dataset captures the characteristics of original digitized historical film reels and produced documentary films and provides researchers a fundamental base to work on automated historical film analysis tools. As representative frame, the center frame of the shot has been chosen.

The validity of the selection of the center frame to represent a shot is done, on the one hand by a manual assessment, and a conducted user study on the other. This comprehensive user study involves experts as well as non-experts from the film domain. The results show that the shot type can be classified by using the single center frame of the shot. Additionally, classification models based on Convolutional Neural Networks (CNN) are presented to provide a baseline for the automated classification. The **contribution of this paper** is summarized as follows:

- We provide a *novel dataset, called Historical Film Shot Dataset (HistShot)* in order to promote research on automated preservation of large historical film archives.
- *A comprehensive user study* including film experts and non-experts is given to approve the validity of our dataset, especially the representation of a shot using the center frame. Furthermore, a *quantitative assessment* is proposed by evaluating CNN baseline models.
- For reproducibility the *source code* as well as *the dataset* are published on Github[1] and Zenodo[2].

[1]https://github.com/dahe-cvl/ICPRAM2022_histshotV1
[2]https://doi.org/10.5281/zenodo.5770202

This paper is structured as follows: In Section 2 a detailed outlook about state-of-the-art movie datasets is demonstrated, and the challenges and drawbacks are illustrated. The proposed Historical Film Shot Dataset (HistShot) is presented and discussed in detail in section 3. Moreover the validity is evaluated in Section 4. Section 5 concludes our investigation with a summary and outlook for future investigations.

2 RELATED WORK

In the last years, a lot of datasets were published to promote research on different video analysis tasks such as action recognition (Gu et al., 2018; Marszalek et al., 2009) or video-text captioning (Miech et al., 2019; Tapaswi et al., 2016). All these datasets include short video clips which are prepared to work on specific tasks. Only a few datasets exist which capture the challenges of professionally produced modern films such as Pirates of the Caribbean or Titanic (Huang et al., 2020; Savardi et al., 2021; Awad et al., 2021).

The dataset **Cinescale** published by (Savardi et al., 2021; Benini et al., 2019; Svanera et al., 2019) was developed in order to provide the research community a basic dataset to work on frame-based methods for detecting shot types in modern movies. This dataset includes 792k frames gathered from 124 different art movies. Each frame corresponds to one of the class categories: Extreme Close Up (ECU), Close Up (CU), Medium Close Up (MCU), Medium Shot (MS), Medium Long Shot (MLS), Long Shot (LS), Extreme Long Shot (ELS), Foreground Shot (FS), and Insert Shots (IS). The whole dataset was annotated by two different annotators and finally, a third one made final corrections. In order to evaluate the dataset, they explore different CNN architectures (Alexnet, VGG16, and GoogleNet) (Savardi et al., 2018). The trained network, with a VGG16 backbone architecture, achieves an accuracy of about 94%. On the official database website[3] the authors point out that the latest model using a Densenet backbone architecture increases the accuracy by about 3%.

A further film database, published by (Vicol et al., 2017), is called **MovieGraphs**. This dataset includes 7637 clips showing human-centric situations extracted from 51 different movies. Each clip is manually labeled with a graph, a situation label, a scene label, and a natural language description. This dataset provides an interesting base for video analysis and deeper research on abstract scene understanding in

[3]https://cinescale.github.io/ - last visited: 2021/10/22

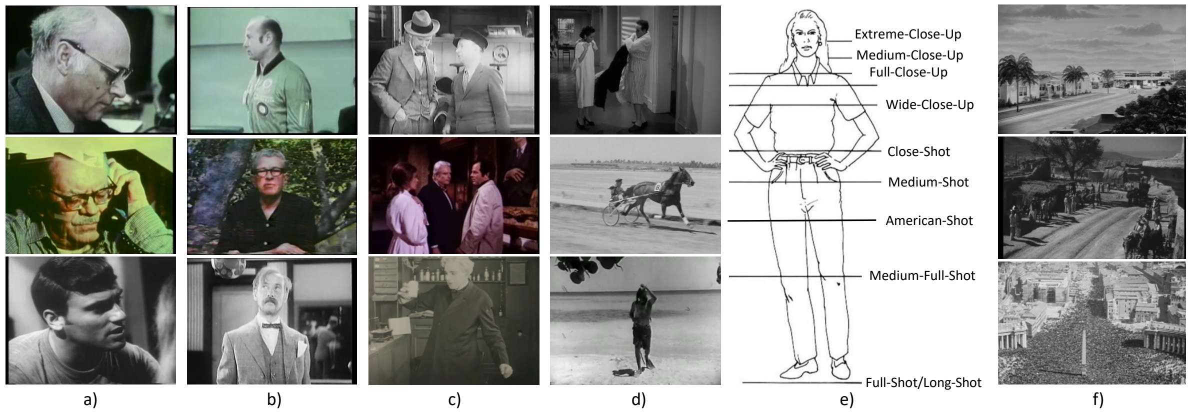

Figure 2: This figure gives an overview of examples of common shot types. (a) Close-Up, (CU), (b) Medium Close Shot, (c) Medium Shot (MS), (d) Long Shot (LS), (e) Schematic overview of shot types and borders, (f) Extreme Long Shot (ELS). Courtesy of (Kahle, 1996).

complex synthetically generated scenes. The graph gives an insight into the characters and their relationships. In order to annotate that the large number of video clips, a web-based annotation tool was developed, and a group of introduced annotators was hired from the platform Upwork[4]. While the dataset describes complex situations with graph representations, no information about fundamental cinematographic techniques such as the shot boundaries or the shot type used to record a scene is available.

Movienet is a further recently published dataset (Huang et al., 2020; Rao et al., 2020). This massive dataset includes annotations for over 1000 movies, 60000 trailers, and 3.9 million photos. Additionally, text-based metadata for each movie is collected. The trailers and photos are gathered from Youtube linked to the corresponding entity in the TMDB[5] and IMDB[6] dataset. Moreover, metadata information about the content is collected from those platforms. The authors provide bounding box annotations for objects, scene segmentation masks, and text-based action and place tags. Moreover, cinematographic settings such as the shot boundaries, camera movements, lighting, and shot types are included and manually labeled. The focus of the shot types is on Extreme-Close-Up, Medium-Shot, Full-Shot, Long-Shot, and Close-Up (Rao et al., 2020). Finally, about 65k cinematographic shots from over 1000 movies are annotated and partly available to the research community. They evaluated the MovieNet dataset in different ways. In order to provide a benchmark for shot type classification, they explore their dataset on three different approaches. I3D (Carreira and Zisserman, 2017), TSN (Wang et al., 2016) and R3Net (Deng et al., 2018) are adopted to classify the shot type of a sequence. The results demonstrate results of over 87.5% accuracy and demonstrate the validity of their dataset.

To our best knowledge, a film dataset containing annotated cinematographic settings in historical films is not publicly available. Their exist a various number of large historical film archives such as **Ephemeral Films (EFilms)**(Zechner, 2015) or **IMediaCities**(Zechner and Loebenstein, 2016). Those platforms include a rich collection of historical films recorded in the last 100 years. Film historians or archivists are able to work with those collections and use the platform to find and edit specific content, such as the shot type used to record a specific situation. A more recent project is the **Visual History of the Holocaust Media Management and Search Infrastructure (VHH-MMSI)**(Zechner and Loebenstein, 2019). It will allow users to work with original footage related to the Second World War and National Socialism with computational-assisted annotation tools.

All previously described public film-related datasets have their strengths and weaknesses. For example, the Movienet dataset includes an enormous number of annotated movies. However, the dataset is only partly available to research communities outside the Asian region. All annotations of the movies are available, but the exact movie versions cannot be published due to **copyright constraints**. Researchers have to find the movies on other platforms and/or often have to pay for them. In general, copyright constraints are one major challenge to publishing a new dataset to the computer vision community. Modern productions or historical films contained in film archives such as (Kahle, 1996), (Government, 1993) or (Government, 1934) do not allow free use of their films. Moreover, the manual annotation process of a

[4]https://www.upwork.com/ - last visited: 2021/10/22
[5]https://www.themoviedb.org/ - last visit: 2021/10/22
[6]https://www.imdb.com/ - last visit: 2021/10/22

usable dataset is very **time consuming** and **cost intensive**. A further drawback is that most film datasets such as MovieGraphs or Cinescale include frame-based or short sequence-based annotations of modern film productions such as Forest Gump or Pirates of the Caribbean. All included movies count to the group of feature films. Less focus is taken on **historical film documentations** and **original digitized footage** (Helm and Kampel, 2019a; Helm and Kampel, 2019b). These films demonstrate specific characteristics such as damages in the film reels, the quality, and the camera techniques used to record a situation. Therefore, the **usability** of public datasets to work on automated film analysis tools such as a shot type classifier is limited.

3 HISTORICAL FILM SHOT DATASET (HistShot)

The Historical Film Shot Dataset (Histshot) consists of 1885 images showing six different types of camera shot types: Close-Up (CU), Medium-Shot (MS), Long-Shot (LS), Extreme-Long-Shot (ELS), Intertitle (I) and Not Available/Not Clear (NA). Those samples are selected from 57 different film recordings of the U.S. National Archives and Records Administration (NARA) (Government, 1934), Film Archive of the Estonian Film Institute (EFA)[7] and the Library of Congress (LoC)[8],[9] related to the Second World War and the time of the Holocaust. During the ongoing research project Visual History of the Holocaust (VHH), the original film reels are digitized and imported to the film archive VHH-MMSI. The film reels are recorded from different cameramen of the U.S. and Soviet forces and demonstrate the liberation phase of Nazi Concentration Camps such as Dachau, Mauthausen, or Bergen-Belsen. Moreover, the recordings visualize the daily-life situations of soldiers and civilians. Figure 3 demonstrates examples from the proposed HistShot dataset.

Each class category in the dataset includes, finally, about 314 frames (average). These images are gathered by using a specified collection process. First, we split each film into its basic shots by following the strategy published by (Helm and Kampel, 2019a). In the next step, the shots are classified into the defined shot type categories by using the pre-trained models of (Helm and Kampel, 2019b). The results are manually corrected by experts of the VHH project consortium. After this process, the most representative frame from each shot has to be extracted. Therefore, a manual assessment points out that the center frame is a valid choice. This is additionally shown by the results of the user study (see Section 4). After extracting all center frames of the individual shots, a first version of the dataset, including about 6000 frames, is generated. As the last step, false predictions and highly similar frames are dropped during a manual filtering process to provide a dataset with a broad spectrum of data related to the content and the film sources. Finally, the published dataset contains 1885 frames, distributed to six class categories with about 314 frames (average). All frames are published with the original resolution of the digitized footage with 1440 by 1080 pixels and the 3 RGB channels. Details about the dataset distribution are visualized in Table 1. More details about the film sources are given at the Zenodo dataset page[10].

Table 1: Details about the proposed Historical Film Shot Dataset (HistShot).

HistShot	All	CU	MS	LS	ELS	I	NA
LoC-EFA	808	76	233	188	170	43	98
NARA	1077	207	209	212	206	181	62
N-Samples	**1885**	**283**	**442**	**400**	**376**	**224**	**160**
Num-LoC-EFA	6	6	6	6	6	6	6
Num-NARA	51	39	39	38	36	50	24
N-Films	**57**	**45**	**45**	**44**	**42**	**56**	**30**

tinyHistShot Dataset: To conduct a user study, a subset of the dataset has been chosen for manual classification. For each category, 20 images have been randomly selected. The small subset is referenced as tinyHistShot DS and has been presented to experts and non-experts. The results of the manual classification are presented in Section 4 and show that the center frame of each shot can be used for classification.

[7]https://www.filmi.ee/ - last visit: 2021/10/28

[8]https://lccn.loc.gov/91796865, Collection: *World War II color footage*, Director: *George Stevens*, between 1943-1945, United States. - last visit: 2021/10/28

[9]https://lccn.loc.gov/91483179, Collection: *World War II black and white footage/Special Coverage Motion Picture Unit - U.S. Army Signal Corps*, Director: *George Stevens*, between 1944-1945, United States. - last visit: 2021/10/28

[10]https://doi.org/10.5281/zenodo.5770202

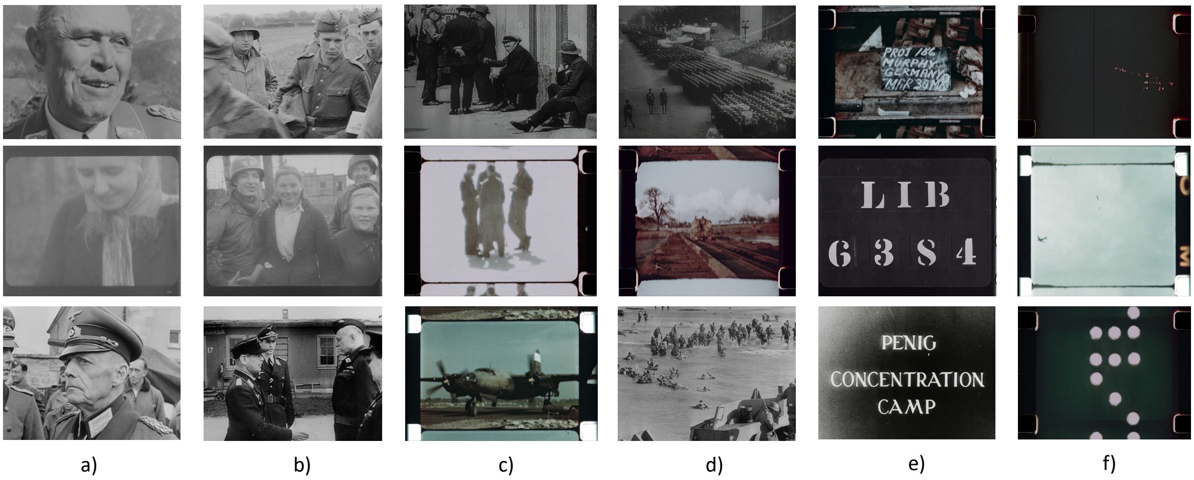

Figure 3: This Figure demonstrates samples of the proposed Historical Film Shot Dataset. (a) Close-Up (CU), (b) Medium-Shot (MS), (c) Long-Shot (LS), Extreme-Long-Shot (ELS), Intertitle (I) and Not Available/Not Clear (NA).

4 EVALUATION

CNN-Baseline: To evaluate our proposed HistShot dataset, we conduct two experiments based on state-of-the-art shot type classification strategies (Savardi et al., 2018; Cherif et al., 2007). Common CNN architectures for solving image classification tasks such as Resnet50 and VGG16 are used in this investigation. Those models are used with the pre-trained ImageNet weights and adapted to classify the introduced shot type categories. In *Experiment 1*, only the frames extracted from the NARA films are involved in the training procedure. The trained models are tested on the frames extracted from the LoC-EFA library and the randomly selected subset for the user study. In *Experiment 2* the entire HistShot dataset (including NARA as well as Loc-EFA frames) is used to train both models. The final test is done on an independently generated testset. This testset includes films imported to VHH-MMSI such as "Schindlers Liste (by Steven Spielberg), Judgement at Nuremberg (by Stanley Kramer) or Die Befreiung von Ausschwitz (by Irmgard von zur Muehlen). This test set includes 10587 samples gathered from 26 different films. Those productions are mainly new visual representations of the time of National-Socialism and count as feature films. This dataset is currently not available to the public due to copyright constraints. However, the trained models in this experiment are also evaluated on the user-study dataset. Table 2 summarizes all quantitative evaluations of the proposed HistShot dataset, and Figure 4 demonstrates the training history of the models based on our dataset and that the trained models are able to generalize well on unseen film frames from different domains.

Table 2: Summary of CNN baseline results. (VM) VHH-MMSI, (HS) HistShot, (tinyHS) tinyHistShot Dataset.

Exp+Backbone+Train+Test	Acc	P	R	F1	N
1+Vgg16+NARA+tinyHS	0,81	0,82	0,81	0,80	120
1+Resnet50+NARA+tinyHS	0,77	0,78	0,77	0,76	120
1+Vgg16+NARA+LoC-EFA	0,81	0,82	0,84	0,82	808
1+Resnet50+NARA+LoC-EFA	0,80	0,81	0,83	0,81	808
2+Vgg16+HS+tinyHS	0,83	0,83	0,83	0,82	120
2+Resnet50+HS+tinyHS	0,78	0,78	0,78	0,77	120
2+Vgg16+HS+VM	0,76	0,70	0,74	0,71	10857
2+Resnet50+HS+VM	0,73	0,67	0,72	0,69	10857

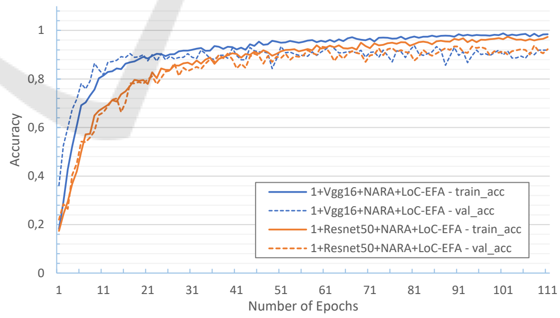

Figure 4: Demonstration of the training history of Experiment $1+Vgg16+NARA+LoC-EFA$ and $1+Resnet50+NARA+LoC-EFA$.

User Study: To conduct a user study, a subset of the HistShot DS with 20 images per class has been defined (tinyHistShot DS). The images have been shuffled and presented to 20 participants (6 experts and 14 non-experts). The numerical results for all participants, experts, and non-experts are shown in Table 3. There is no significant difference between experts and non-experts, which shows that no additional ex-

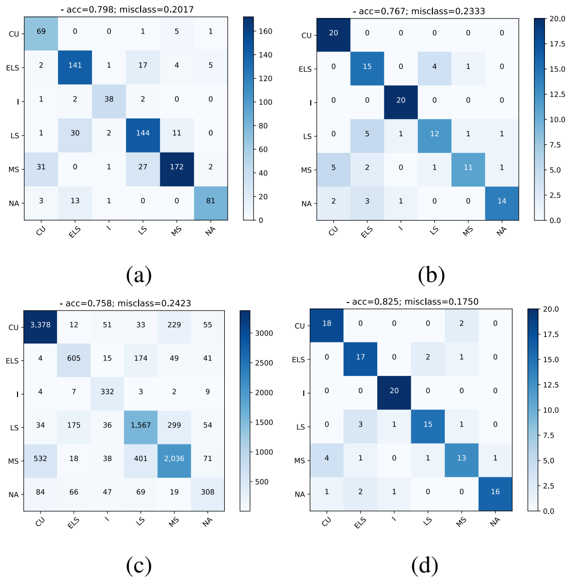

Figure 5: Shot Type classification performance illustrated by the confusion matrices of different baseline experiments. (a) 1+Vgg16+NARA+LoC-EFA, (b) 1+Vgg16+NARA+tinyHS, (c) 2+Vgg16+HS+VM, (d) 2+Vgg16+HS+tinyHS.

pert knowledge is needed for classification. The result of the overall classification is shown in Figure 6. The overall result shows that the center frame is a valid choice to describe a shot with a single frame. The confusion matrix shows that errors occur between the classes Close-Up (CU) and Medium-Shot (MS), Medium-Shot (MS) and Long-Shot (LS), Long-Shot (LS), and Extreme Long Shot (ELS). Additionally, Intertitles (I) can be interpreted as Close-Ups (CU). This shows that there is no fixed measurable boundary between consecutive classes, which are "*loosely defined areas on a continuum of camera-to-subject distance*"[11]. In Figure 7 some examples of challenging scene situations are demonstrated.

Table 3: Summary of the Results of the User Study.

User Study	Acc	P	R	F1
All Participants (20)	0,63	0,66	0,63	0.64
Experts (6)	0,60	0,66	0,60	0,61
Non-Experts (14)	0,63	0,66	0,63	0,64

A more "sharp" transition between the classes can be achieved with machine learning compared to human classification. The baseline methods show a clear improvement between consecutive classes (up to 20% higher F_1-score).

[11]http://www.filmreference.com/encyclopedia/Romantic-Comedy-Yugoslavia/Shots-CLASSIFICATION-OF-SHOTS.html - last visit: 2021/10/28

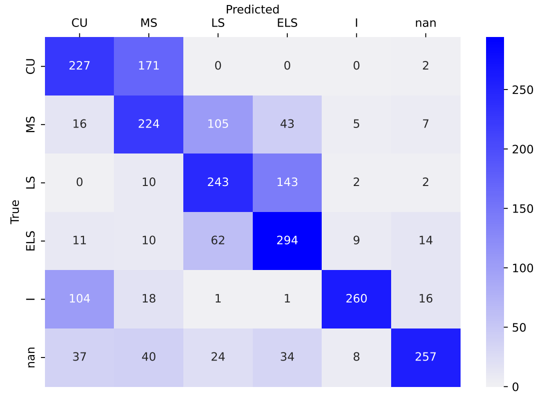

Figure 6: Result of the User Study (Shot type classification annotated by 20 individuals).

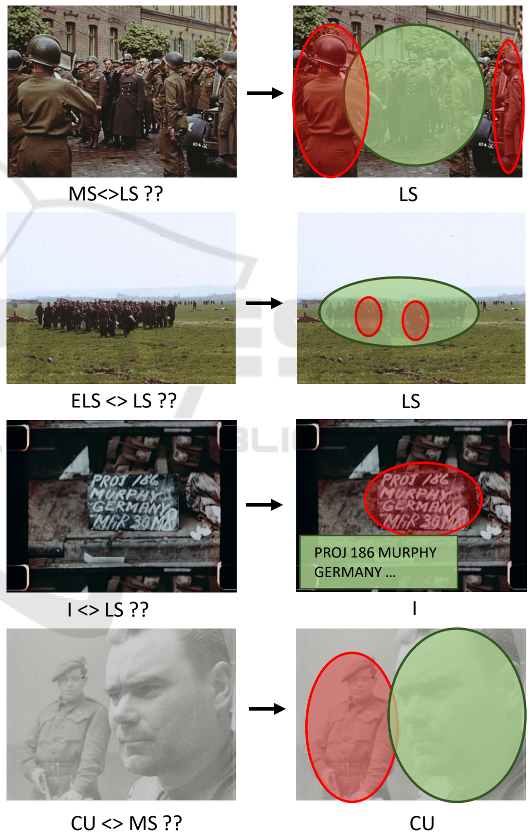

Figure 7: Challenging examples from different shot type categories. The *green bubble* illustrates the most significant area of a frame including the subject of interest whereas the *red bubbles* demonstrate other scene subjects which are not on the focus of the camera man.

5 CONCLUSIONS

A freely available shot type dataset based on historical films (1943-1945) has been presented. Each shot is represented with the center frame, and baseline methods (ResNet50, VGG16) are established. The dataset has a size of 1885 samples and six different shot types. Additionally, a user study has been conducted to compare the results with human classification. Compared to state-of-the-art datasets (e.g., Cinescale, MovieGraphs), the published HistShot dataset focuses on historical documentaries and original digitized film reels. In a follow-up investigation, the dataset will be extended with additional cinematographic annotations such as shot boundaries, shot-based shot types, and camera movements. Moreover, the dataset will include exclusive original digitized footage related to the Second World War (about 100 films).

ACKNOWLEDGEMENTS

Visual History of the Holocaust: Rethinking Curation in the Digital Age (Zechner and Loebenstein, 2019). This project has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement 822670. Special thanks to all participants (film experts and non-experts) of the survey.

REFERENCES

Awad, G., Butt, A. A., Curtis, K., Fiscus, J. G., Godil, A., Lee, Y., Delgado, A., Zhang, J., Godard, E., Chocot, B., Diduch, L. L., Liu, J., Smeaton, A. F., Graham, Y., Jones, G. J. F., Kraaij, W., and Quénot, G. (2021). TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains. *CoRR*, abs/2104.13473.

Benini, S., Savardi, M., Bálint, K., Kovács, A. B., and Signoroni, A. (2019). On the influence of shot scale on film mood and narrative engagement in film viewers. *IEEE Transactions on Affective Computing*, pages 1–1.

Benini, S., Svanera, M., Adami, N., Leonardi, R., and Kovács, A. B. (2016). Shot scale distribution in art films. *Multimedia Tools and Applications*, 75(23):16499–16527.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, Los Alamitos, CA, USA. IEEE Computer Society.

Cherif, I., Solachidis, V., and Pitas, I. (2007). Shot type identification of movie content. In *2007 9th International Symposium on Signal Processing and Its Applications*, pages 1–4, Sharjah, United Arab Emirates. IEEE.

Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., and Heng, P.-A. (2018). R^{3}net: Recurrent residual refinement network for saliency detection. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 684–690. International Joint Conferences on Artificial Intelligence Organization.

Flückiger, B., Pfluger, D., Trumpy, G., Aydin, T., and Smolic, A. (2018). Film material-scanner interaction. Technical report, University of Zurich, Zurich.

Fossati, G. (2018). *From Grain to Pixel - The Archival Life of Film in Transition*. Amsterdam University Press, Amsterdam.

Fossati, G. and van den Oever, A. (2016). *Exposing the Film Apparatus*. Amsterdam University Press, Amsterdam.

Government, U. S. (1934). The U.S. National Archives and Records Administration. https://www.archives.gov/. [Online; last accessed 31.05.2021].

Government, U. S. (1993). United States Holocaust Memorial Museum. https://www.ushmm.org/. [Online; last accessed 31.05.2021].

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., and Malik, J. (2018). Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Helm, D. and Kampel, M. (2019a). Shot boundary detection for automatic video analysis of historical films. In Cristani, M., Prati, A., Lanz, O., Messelodi, S., and Sebe, N., editors, *New Trends in Image Analysis and Processing – ICIAP 2019*, pages 137–147, Cham. Springer International Publishing.

Helm, D. and Kampel, M. (2019b). Video Shot Analysis for Digital Curation and Preservation of Historical Films. In Rizvic, S. and Rodriguez Echavarria, K., editors, *Eurographics Workshop on Graphics and Cultural Heritage*. The Eurographics Association.

Huang, Q., Xiong, Y., Rao, A., Wang, J., and Lin, D. (2020). Movienet: A holistic dataset for movie understanding. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12349 LNCS:709–727.

Kahle, B. (1996). Internet archive. https://archive.org/. [Online; last accessed 2020/11/09].

Luca, C., Sergio, B., and Riccardo, L. (2013). Classifying cinematographic shot types. *Multimedia Tools and Applications*, 62(1):51–73.

Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Rao, A., Wang, J., Xu, L., Jiang, X., Huang, Q., Zhou, B., and Lin, D. (2020). A unified framework for shot type classification based on subject centric lens. In *The European Conference on Computer Vision (ECCV)*.

Savardi, M., Kovács, A. B., Signoroni, A., and Benini, S. (2021). CineScale: A dataset of cinematic shot scale in movies. *Data in Brief*, 36:107002.

Savardi, M., Signoroni, A., Migliorati, P., and Benini, S. (2018). Shot scale analysis in movies by convolutional neural networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2620–2624.

Svanera, M., Savardi, M., Signoroni, A., Kovács, A. B., and Benini, S. (2019). Who is the film's director? authorship recognition based on shot features. *IEEE MultiMedia*, 26(4):43–54.

Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., and Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vicol, P., Tapaswi, M., Castrejon, L., and Fidler, S. (2017). MovieGraphs: Towards Understanding Human-Centric Situations from Videos. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 20–36, Cham. Springer International Publishing.

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

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

Zechner, I. and Loebenstein, M. (2019). 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/. [Online; last accessed 31.08.2020].