Graph-based Shot Type Classification in Large Historical Film Archives

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 Classification, Cultural Heritage, Graph Neural Network.

Abstract: To analyze films and documentaries (indexing, content understanding), a shot type classification is needed. State-of-the-art approaches use traditional CNN-based methods, which need large datasets for training (CineScale with 792000 frames or MovieShots with 46K shots). To overcome this problem, a Graph-based Shot TypeClassifier (GSTC) is proposed, which is able to classify shots into the following types: Extreme-Long-Shot (ELS), Long-Shot (LS), Medium-Shot (MS), Close-Up (CU), Intertitle (I), and Not Available/Not Clear (NA). The methodology is evaluated on standard datasets as well as a new published dataset: HistShotDS-Ext, including 25000 frames. The proposed Graph-based Shot Type Classifier reaches a classification accuracy of 86%.

1 INTRODUCTION

Film shot retrieval in large film archives is an ongoing problem for film archivists, historians, or computer vision researchers (Zechner and Loebenstein, 2019; Helm et al., 2020). Different challenges such as the large available number of different recordings or the broad spectrum of content (documentary/feature films or amateur/professional films) make it not trivial to find specific situations in those collections. However, retrieving specific recordings is one crucial objective of historians or film archivists to provide new visual representations of specific domains of the humans' cultural history such as the time of the Second World War (Zechner, 2015; Zechner and Loebenstein, 2016). A fundamental process for experts is to search and annotate interesting situations manually, which is time-consuming and cost-intensive. Therefore, automated tools are needed to provide experts efficient and innovative ways for sustainable preservation of large historical film archives. One fundamental step for automated film shot retrieval is to understand basic cinematographic settings used to record a situation such as Shot Type Classification (STC) or Camera Movements Classification (CMC).

This work focuses on efficient classification of shot types (or shot sizes) in large film archives. Shot types are used to give a film shot a specific characteristic. This setting is a kind of representation of the distance between the subject of interest (e.g., a person) and the camera lens. There are several definitions of shot type categories, and there is no defined unique standard. However, one common definition of shot type categories is: Extreme-Close-Up (ECU), Medium-Close-Up (MCU), Full-Close-Up (FCU), Wide-Close-Up (WCU), Close-Shot (CS), Medium-Shot (MS), American Shot (AS), Medium-Full-Shot (MFS), Full-Shot (FS) and Extreme-Long-Shot (ELS). A schematic illustration of the shot type borders is demonstrated in Figure 1b. Each type is used to give a professionally recorded shot a specific characteristic. For example, an FCU is used to point out strong emotions, whereas an ELS lets the observer dive into the depth of a scene-setting (panorama view). Some examples of historical films as well as of modern film productions are given in Figure 1a. Current state-of-the-art shows the applicability of graph representation learning in different research domains such as scene graph generation (Marino et al., 2016) or graph-based image retrieval (Yoon et al., 2020). Traditional CNNs need an enormous number of training samples and take a lot of training duration. They try to learn meaningful features from each image in order to generalize to unseen data (Russakovsky et al., 2015; Simonyan and

[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.
Graph-based Shot Type Classification in Large Historical Film Archives.
DOI: 10.5220/0010905800003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 991-998
ISBN: 978-989-758-555-5; ISSN: 2184-4321

Figure 1: Background about the cinematographic film setting: *Shot-Type* (Shot-Size). (a) Some examples of different shot-types (Kahle, 1996) (b) Schematic visualization of different shot type categories with respect to a human actor.

Zisserman, 2015). The idea of graph neural networks is to represent a complex data structure such as a large image database or film archive as a graph (Misraa et al., 2020). One data point in a graph is represented as a so-called *node*. *edges* are used to concatenate related nodes. All nodes and edges in a graph can have individual *attributes*. The most significant benefit of graph representation learning for image classification is that the neighborhood of an image (Node) in a graph is considered to learn an aggregation function which is finally able to classify the target node into a specific category (Wu et al., 2019).

In this paper, a novel approach to classify film shot types in large historical film archives is presented. The pipeline consists of three stages: *Shot-based Keyframe Selection & Feature Extraction*, *Graph Generator* and the *Graph-based Shot Type Classifier (GSTC)* and is able to classify frames into one of the six shot type categories. The best classification result (Accuracy=86%) is reached by using Resnet152 features in combination with a Graph Attention Network (GAT). The GSTC is trained and evaluated with a self-generated dataset based on the Historical Film Shot Dataset (Helm et al., 2022) called *HistShotDS-Ext* and includes about 35000 samples of about 130 different films. The results point out the effectiveness of the proposed graph-based pipeline compared to traditional Convolutional Neural Network classifiers.

The **contribution of this paper** is summarized as follows:

- We provide a *novel approach based on Graph Neural Networks* to classify film shot types into the categories: Extreme-Long-Shot (ELS), Long-Shot (LS), Medium-Shot (MS), Close-Up (CU), Intertitle (I), and Not Available/Not Clear (NA). Moreover, a *baseline for graph-based node classification* in large film archives is given to the research community.
- An extension to the Historical Film Shot Dataset, called *HistShotDS-Ext* is provided to promote further research on historical film analysis. The dataset includes about 35000 annotations from about 130 different films.
- For reproducibility, the *source code*, as well as *the annotations and film sources*, are published on Github[1].

The rest of this paper is organized as follows: The state-of-the-art and related work is described in Section 2. A detailed description of the methodology is presented in Section 3. The experimental setup and details about the dataset used in this investigation are described in Section 4. All results are presented and discussed in Section 5. Finally, the paper concludes with Section 6.

2 RELATED WORK

Image classification tasks are mainly solved by using standard CNN architectures such as the Resnet50/152 (Sangeetha and Prasad, 2006), or VGG16 (Simonyan and Zisserman, 2015) with a specified classification header to predict customized class categories. Results over 90% accuracy are reached on different datasets such as Cifar10 (Krizhevsky, 2009), ImageNet (Russakovsky et al., 2015) or MS COCO (Lin et al., 2014). The major drawback is that those models need a lot of pre-labeled training data, e.g., in the size of 50000 images (Cifar10) or up to 1.2M images (ImageNet). However, current state-of-the-art methods have established graph representation learning (Zhang et al., 2019), which is used in different domains such as scene understanding (Liang et al., 2020), knowledge graph learning (Marino et al., 2016), image classification (Avelar et al., 2020; Nikolentzos et al., 2021) and image similarity measures (Yoon et al., 2020) and show superior results. The authors of (Marino et al., 2016) propose a method for multi-label image classification by using graph-based representations of images. (Long et al., 2021) propose a Graph Attention Network and (Nikolentzos et al., 2021) follow a similar approach using a kings graph and coarsened graph.

Work on classifying cinematographic settings, such as shot types, is presented by (Savardi et al., 2021; Rao et al., 2020) or (Savardi et al., 2018). All authors focus on traditional CNN-based techniques such as GoogleLeNet, AlexNet, or VGG16 and use manually labeled datasets with up to 792000 image frames (CineScale) or 46K shots (MovieShots). (Vretos et al., 2012) present 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 out of

[1]https://github.com/dahe-cvl/VISAPP2022_GSTC

seven class categories: ECU (Extreme Close Up), CU (Close Up), MCU (Medium Close Up), MS (Medium Shot), MLS (Medium Long Shot), LS (Long Shot) and ELS (Extreme Long Shot). Therefore, an SVM classifier is applied to the extracted features. The major drawback of this solution is that it assumes that in each image, you see at least one person otherwise, no classification is possible.

Contrary to standard approaches in classifying cinematographic settings, this paper proposes a Graph-based Shot Type Classifier (GSTC). Based on discussions with film experts and archivists, the shot categories (ELS, LS, MS, CU, I, NA) have been chosen as the most descriptive shot types for analysis.

3 METHODOLOGY

The proposed approach in this paper consists of several stages: Shot-based Keyframe Selection & Feature Extraction, Graph Generator and Graph-based Shot Type Classifier (GSTC). Figure 2 illustrates a schematic overview of the entire pipeline.

3.1 Shot-based Keyframe Selection & Feature Extraction

The proposed approach classifies individual frames corresponding to a film shot. Thus, the first step is to split a given film in a database into its shots. We use a deep learning-based technique provided by (Helm and Kampel, 2019) to detect hard cuts in given films, and as a next step, select a representative keyframe of each individual shot. This keyframe should represent the most significant information of the entire sequence. A manual evaluation of individual shots demonstrates a broad diversity in terms of recording technique or the content. For example, a shot can be recorded using different camera movements such as a Pan or a Tilt. It can be observed that shots often start with a mixed camera movement (pan-tilt) until the target object (e.g., a person) is in the cameras' focus. Furthermore, recordings often show highly dynamic situations, e.g., many objects (e.g., vehicles, persons, etc.) move through a scene. It is not clear which object is the subject of interest during the shot. However, it has been shown that the center frame is a valid choice for shot classification (Helm et al., 2022) since it represents the most significant content of a shot recording and is therefore used in the proposed evaluation.

After selecting valid keyframes (center frames) from individual shots, we need to extract meaningful visual image features as a base for our proposed graph-based classifier. Therefore, different pre-trained backbone Convolutional Neural Networks (CNN) is used to get a feature vector. The focus in this investigation is on Resnet152 and Resnet50, which are already pre-trained with the ImageNet dataset (Russakovsky et al., 2015). The last layer of the pre-trained model is dropped, and the output of the Average-Pooling-Layer represents the 2048 dimensional feature vector corresponding to a specific keyframe (see Figure 2).

3.2 Graph Generator

The next stage in the proposed pipeline is the Graph Generator. The strategy by (Misraa et al., 2020) has been used where each node in the graph represents one keyframe corresponding to a specific shot. Furthermore, each node is described with a node feature vector. The feature vector of the shot-based keyframe selection and feature extraction stage is assigned to the corresponding node in the graph in contrast to (Misraa et al., 2020). The k-Nearest-Neighbors Graph (kNN-Graph) is used to find edges between individual nodes. The kNN-Graph finds the k nearest neighbors of a target node based on the given node feature vectors. As distance measure, the euclidean distance is used, and the parameter k is selected empirically. The output of the kNN-Graph is an adjacency matrix that represents the binary node connections of the entire graph. Additionally, the euclidean distances between the node feature vectors are used as edge attributes. Figure 2 illustrates a simplified graph including node features and 1-dimensional edge attributes. The adjacency matrix includes self-loops and is represented as an undirected graph.

3.3 Graph-based Shot-Type-Classifier (GSTC)

The last stage in the pipeline is the Graph-based Shot-Type-Classifier (GSTC) based on a Graph Neural Network architecture. In this work, experiments with different layer architectures such as the standard Graph Convolutional Neural Network (GCN) (Kipf and Welling, 2016), Graph SAmple and aggreGatE (GraphSAGE) (Hamilton et al., 2017), and Graph Attention Network (GAT)(Veličković et al., 2018) are evaluated. The model consists of three layers and is trained to solve a node classification task. The input of GSTC is a sub-graph sampled from the entire database graph. The graph sampling strategy used is introduced in the GraphSage paper (Hamilton et al., 2017). It demonstrates better generalization for inductive representation learning and an effi-

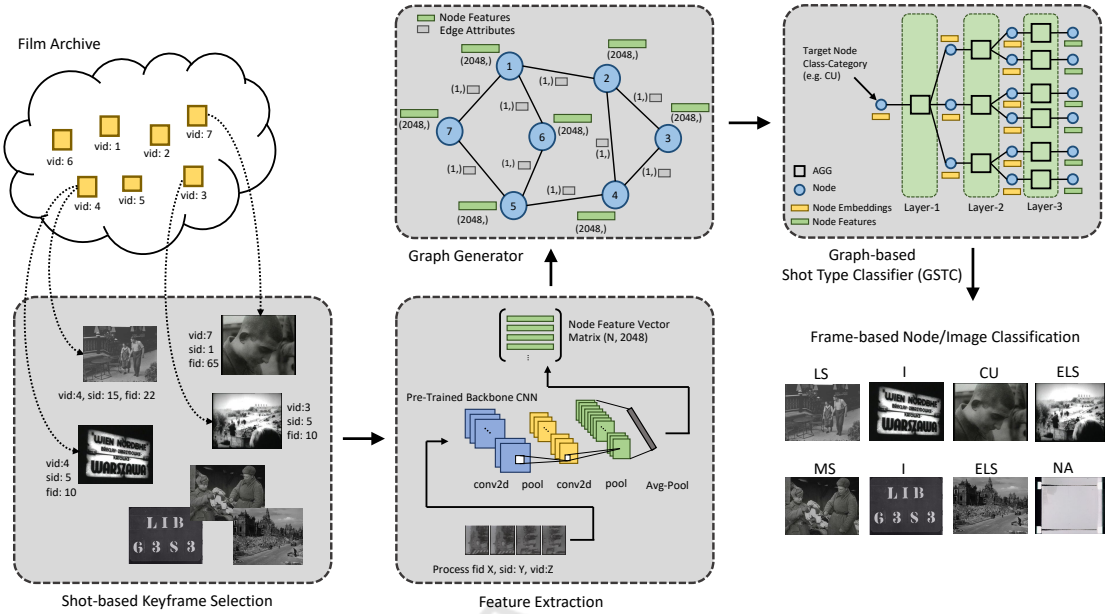


Figure 2: Schematic overview of the Graph-based Shot-Type Classifier using a Graph Neural Network for node classification.

cient way to sample data from large graphs (Hamilton et al., 2017). After the sampled graph is fed into the GNN, the final node embedding for the target node is calculated by aggregating the nodes' neighborhood. After each layer except the output layer, a Rectified Linear Unit (ReLU activation) and a Dropout Layer with a probability of 0.5 are applied. Finally, the target node embedding is used to classify the corresponding node into one out of the six possible classes: Extreme-Long-Shot (ELS), Long-Shot (LS), Medium-Shot (MS), Close-Up (CU), Intertitle (I), and Not Available (NA). A schematic overview of the model architecture is given in Figure 3.

4 EXPERIMENTAL SETUP

The following subsections describe the datasets used, training details, and a general description of the experiments.

4.1 Dataset

All experiments are based on the **Historical Film Shot Dataset (HistShotDS)** (Helm et al., 2022). This dataset includes 1885 frames extracted from 57 original digitized film reels stored in the U.S. National Archives and Records Administration (NARA) (Government, 1934), Film Archive of the Estonian Film Institute (EFA)[2] and the Library of Congress (LoC)[3,4]. In order to simulate a large film archive the HistshotDS is extended with further frames extracted from different feature and documentary films stored in the Internet-Archive (*Silent Films Collection*, *The Video-Cellar Collection*) (Kahle, 1996), EFilms (Zechner, 2015) and Imediacities(Zechner and Loebenstein, 2016). Finally, the entire training dataset (**HistshotDS-Ext**) includes 25000 frames extracted from about 100 different historical films mainly related to the Second World War and the time of National Socialism. The films show real-world situations, including many different objects such as persons or vehicles in a highly dynamic environment. The 25000 frames are gathered by extracting the center frames of each shot because the center part of a shot holds the most significant information (Helm et al., 2022). This set of images is used for the training and validation procedure. A separate dataset is generated for testing, including 10857 frames extracted from 30 films not included in the training set. Due to copyright constraints, **HistshotDS-Ext** can not be published. However, the information, including film title, origin, and frame numbers, is made available on Github to provide a reproducibility of the proposed dataset. Moreover, to get a basic understand-

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

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

[4]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

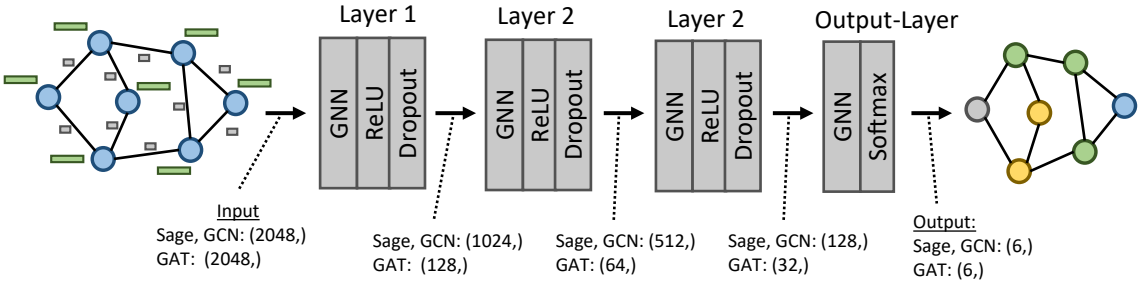


Figure 3: A schematic overview of the Graph-based Shot-Type Classifier using Graph Neural Network for node classification. In this paper the focus is on GraphSage, GCN and GAT.

ing of how the proposed Graph-based STC approach works on state-of-the-art datasets, the benchmark set Movienet (Huang et al., 2020) is evaluated. The Movienet-Version1 published by (Rao et al., 2020) includes about 22401 frames from different film trailers corresponding to the categories: Extreme-Close-Shot (ECS), Close-Shot (CS), Medium-Shot (MS), Full-Shot (FS), and Long-Shot (LS). This dataset is split into a training set (17920) and a separate test set (4481). All proposed results are evaluated on the test set. Moreover, standard metrics such as Precision, Recall, F1-Score, and Accuracy are used to evaluate the classification task.

4.2 Training Details & Inference Mode

Training Details: For all experiments, the Adam optimizer in combination with the Cross-Entropy loss function is used. Furthermore, a batch size of 128 and a maximum training time of 500 epochs are configured. In order to avoid overfitting, early-stopping and learning rate decay (patience of 25 epochs) are implemented. Moreover, Dropout (probability: 0.5) and weight decay (GCN: 0.002, GraphSage: 0.005 - GAT: 0.0005) are applied in training mode. More details about the implementation can be found in the GitHub Repository.

Inference Mode: In inference mode, a new data sample is added to the entire graph by calculating the k-nearest-neighbors. Finally, the class category is predicted by calculating the new node embedding of the corresponding node. In **Test Mode**, all test samples (10587) are induced in the entire database graph by using the same strategy as previously described. The test evaluation is done by calculating all new node embeddings and the classification performance.

4.3 Experiments

The following experiments have been conducted:

Classification Performance: In order to evaluate the classification performance of shot types by using graph representations, several combinations of CNN backbone features and GNNs are defined. In our investigation, the combination of Resnet152 and Resnet50, pre-trained on ImageNet, with Graph Convolutional Networks (GCN), Graph Sample and Aggregate (GraphSage), and Graph Attention Networks (GAT) are evaluated. Additionally, to compare the results of the proposed Graph-based Shot-Type-Classifier (GSTC) with state-of-the-art methods, traditional CNN classifiers (Resnet50 and VGG16) are trained on the HistShotDS-Ext.

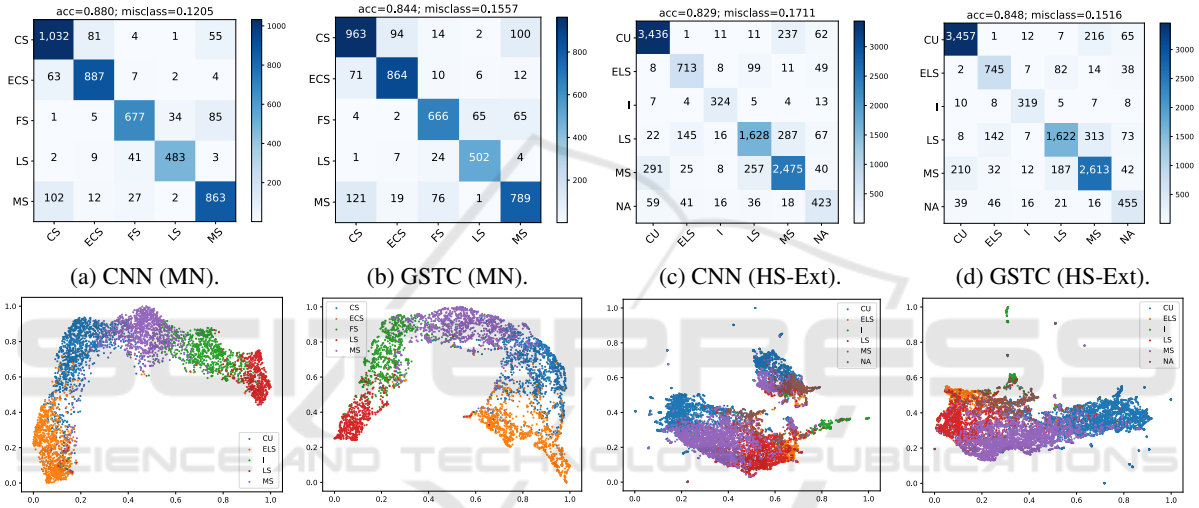
Influence of Neighborhood: A further experiment is the evaluation of different sizes of neighborhoods. The nodes in large graphs are sampled by selecting a target node and the corresponding k-hop neighbor nodes. The assumption is that more selected neighbor nodes of a target node yield higher classification accuracy because more neighborhood information can influence the final model performance.

5 RESULTS

Classification Performance: In Table 1 an overview of the gathered results are demonstrated. All models are pre-trained on the ImageNet dataset and used without further fine-tuning on the target image domain (HistshotDS-Ext). The experiment Resnet152(ImageNet)+GCN with the corresponding neighborhood k:[10, 5, 2] demonstrates an accuracy of 78.52%. The GraphSage model, in combination with the Resnet152 features, shows a significantly better classification performance and reaches an accuracy of 84.31% with a neighborhood of k:[1, 1, 1]. The combination Resnet152+GAT demonstrates the best result with the three-level neighborhood (k: 15, 7, 2). This combination predicts newly added nodes to the entire database graph with an accuracy of 85.6%. In order to compare the results of the proposed approach in this work, the traditional CNN architectures Resnet50 and VGG16 are fine-tuned on the HistshotDS-Ext and accuracies

Table 1: This table visualizes the results of different combinations of CNN backbone features and GNN layers compared to traditional CNN image classifiers.

Experiment	Acc	Prec	Rec	F1	N
CNN-Resnet50(HistShotDS-Ext)	0,83	0,80	0,81	0,80	10857
CNN-VGG16(HistShotDS-Ext)	**0,87**	**0,84**	**0,85**	**0,85**	**10857**
GSTC-Resnet152(IN)+GCN [k: 10, 5, 2]	0,79	0,76	0,77	0,76	10857
GSTC-Resnet152(IN)+Sage [k: 1, 1, 1]	0,84	0,81	0,81	0,81	10857
GSTC-Resnet152(IN)+GAT [k: 15, 7, 2]	**0,86**	**0,83**	**0,84**	**0,83**	**10857**
GSTC-Resnet50(IN)+GCN [k: 10, 5, 2]	0,77	0,73	0,74	0,74	10857
GSTC-Resnet50(IN)+Sage [k: 1, 1, 1]	0,82	0,79	0,79	0,79	10857
GSTC-Resnet50(IN)+GAT [k: 15, 7, 2]	**0,83**	**0,81**	**0,81**	**0,81**	**10857**
GSTC-Resnet50(HistShotDS-Ext)+GAT [k: 15, 7, 2]	**0,84**	**0.81**	**0.83**	**0.81**	**10857**



(a) CNN (MN). (b) GSTC (MN). (c) CNN (HS-Ext). (d) GSTC (HS-Ext).

(e) UMAP - CNN (MN). (f) UMAP - GSTC (MN). (g) UMAP - CNN (HS-Ext). (h) UMAP - GSTC (HS-Ext).

Figure 4: This Figure illustrates the classification performance of individual class categories and the corresponding UMAP plots of extracted feature embeddings: (a)(e) *CNN-Resnet50(MovieNet)*, (b)(f) *Resnet152(ImageNet)+GAT-testset:MovieNet*, (c)(g) *CNN-Resnet50(HistShotDS-Ext)*, (d)(h) *Resnet152(ImageNet)+GAT-testset:HistShotDS-Ext-Test*.

of 86.7% and 82.89% are achieved. Moreover, the model GTSC-Resnet152(ImageNet)+GAT is trained and tested on the state-of-the-art dataset Movienet (Huang et al., 2020) (more details are mentioned in Table 2). The result demonstrates an accuracy of 84%. The final assumption in this work is that the best results are reached by using the fine-tuned Resnet50 on the HistShotDS-Ext in combination with the Graph Attention mechanism. This combination demonstrates an accuracy of 84% with k:[15, 7, 2] and outperforms the traditional CNN-based classifier, Resnet50(HistShotDS-Ext)+GAT (Accuracy=83%). Compared to the CNN-VGG16(HistShotDS-Ext) with the classification performance of 87%, the proposed GSTC-Resnet152(IN)+GAT and GSTC-Resnt50(IN)+GAT demonstrate adequate results (Note that the VGG16/Resnet50 are trained on the HistShotDS-Ext, while the GSTC has no a priori knowledge of the dataset). The evaluation of the benchmark dataset Movienet shows that the fine-tuned Resnet50 with the Movienet dataset in combination with the GAT (GSTC-Resnet50[MN]-GAT) outperforms the traditional CNN-based classifier (CNN-Resnet50[MN]). Figure 4 points out the class category performance (confusion matrices & UMAP plots) of the GSTC approach in comparison to the traditional CNN-based shot type classifier. Moreover, the results on the separately evaluated benchmark dataset Movienet are demonstrated.

Influence of Neighborhood: In this experiment, the influence of different neighborhoods is evaluated. Therefore, the proposed GSTC with the Resnet152(ImageNet)+GAT/SAGE/GCN are trained with different neighborhoods. Figure 5 illustrates the

Table 2: Comparison between traditional CNN-based classifier and our proposed Graph-based Shot-Type-Classifier (GSTC) trained and tested on the Movienet dataset. (MN) Movienet, (IN) ImageNet. (k-neighborhood) [k: 15, 7, 2].

Experiment	Acc	Prec	Rec	F1	N
CNN-Resnet50(MN)	0,88	0,89	0,88	0,88	4482
GSTC-Resnet152(IN)+GAT	0,84	0,85	0,85	0,85	4482
GSTC-Resnet50(MN)+GAT	**0.89**	**0.90**	**0.90**	**0.90**	**4482**

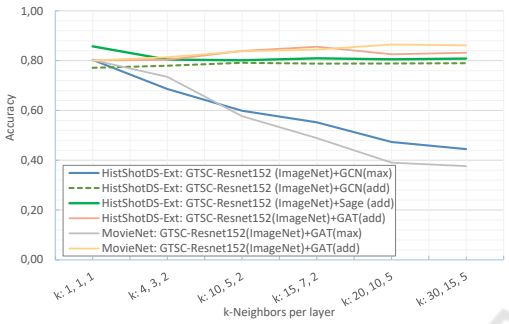


Figure 5: This Figure illustrates the test accuracies by using different k-neighborhood configurations.

accuracies on the HistShotDS-Ext testset on six different configurations by using the *Summation (ADD)* and *Maximum (MAX)* aggregation function. All experiments, using the aggregation *ADD*, demonstrate no significant increase of the accuracy if more neighbors are loaded. Compared to this aggregation, the *MAX* function points out a significant decrease in the classification performance by increasing the target nodes' neighborhood. The reason for this observation is the information loss by using the maximum operator. It drops potential "good" information from the neighborhood, while the ADD function sums up all available feature information. This effect can also be observed in the experiment with the MovieNet dataset. The overall best classification result ($accuracy = 86\%$) is reached with *GSTC-Resnet152(ImageNet)+GAT* in combination with the configuration *k: [15, 7, 2]*.

6 CONCLUSION

A novel baseline approach to classify shot types in large film archives is presented based on modern graph representation learning strategies. Different combinations of visual features extracted from traditional pre-trained CNNs in combination with different Graph Neural Network architectures (GCN, Sage, GAT) are established. The Resnet50/152 architectures fine-tuned on the ImageNet dataset in combination with a Graph Attention Network illustrate accuracies of 83% and 86% on the Historical Film Shot Dataset (Extended) and reach comparable results to traditional CNN-based classifiers (VGG16: 87% & Resnet50: 83%) trained on the proposed dataset. The most impressive result is demonstrated by using a Resnet50 model (pre-trained on the benchmark dataset MovieNet) with the Graph Attention Network (GAT). This combination reaches a classification accuracy of 89% and outperforms the traditional state-of-the-art CNN classifiers (CNN-Resnet50[MovieNet]). This investigation points out the potential power of using graph-based representations for analyzing large film archives.

However, it can be concluded that shot type classification is not a trivial task for humans as well as machines and is not solved (Helm et al., 2022). One reason is the complexity of real-world film scenarios (e.g., a large number of objects, highly dynamically environment), which gives the observer room for interpretation. Computational archival systems need clear conditions to produce accurate results. Different factors in interpreting a recorded frame correctly are mandatory. For example, interpreting the ratio between *the area of the subject and the background* as well as the *depth of a scene* are crucial for classifying shot types. Moreover, a significant question is, *which object(s) is(are) the most interesting one(s) in a scene?*. Therefore, future investigation can be on designing a model that can take all the previously mentioned aspects into account.

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.

REFERENCES

Avelar, P. H., Tavares, A. R., Da Silveira, T. L., Jung, C. R., and Lamb, L. C. (2020). Superpixel Image Classification with Graph Attention Networks. *Proceedings - 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2020*, pages 203–209.

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

Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. *CoRR*, abs/1706.02216.

Helm, D. and Kampel, M. (2019). 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., Kleber, F., and Kampel, M. (2022). HistShot: A Shot Type Dataset based on Historical Documentation during WWII. [will be published soon].

Helm, D., Pointner, B., and Kampel, M. (2020). Frame border detection for digitized historical footage. In Roth, P. M., Steinbauer, G., Fraundorfer, F., Brandstötter, M., and Perko, R., editors, *Proceedings of the Joint Austrian Computer Vision and Robotics Workshop 2020*, pages 114–115, Graz. Verlag der Technischen Universität Graz.

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].

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. pages 32–33.

Liang, Z., Guan, Y., and Rojas, J. (2020). Visual-Semantic Graph Attention Network for Human-Object Interaction Detection. 1.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755.

Long, J., Yan, Z., and Chen, H. (2021). A Graph Neural Network for superpixel image classification. *Journal of Physics: Conference Series*, 1871(1).

Marino, K., Salakhutdinov, R., and Gupta, A. (2016). The more you know: Using knowledge graphs for image classification. *CoRR*, abs/1612.04844.

Misraa, A. K., Kale, A., Aggarwal, P., and Aminian, A. (2020). Multi-modal retrieval using graph neural networks. *CoRR*, abs/2010.01666.

Nikolentzos, G., Thomas, M., Rivera, A. R., and Vazirgiannis, M. (2021). *Image Classification Using Graph-Based Representations and Graph Neural Networks*, volume 944. Springer International Publishing.

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)*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sangeetha, V. and Prasad, K. J. (2006). Syntheses of novel derivatives of 2-acetylfuro[2,3-a]carbazoles, benzo[1,2-b]-1,4-thiazepino[2,3-a]carbazoles and 1-acetyloxycarbazole-2- carbaldehydes. *Indian Journal of Chemistry - Section B Organic and Medicinal Chemistry*, 45(8):1951–1954.

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.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–14.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. (2018). Graph attention networks.

Vretos, N., Tsingalis, I., Nikolaidis, N., and Pitas, I. (2012). Svm-based shot type classification of movie content.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596.

Yoon, S., Kang, W., Jeon, S., Lee, S., Han, C., Park, J., and Kim, E. (2020). Image-to-image retrieval by learning similarity between scene graphs. *CoRR*, abs/2012.14700.

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].

Zhang, S., Tong, H., Xu, J., and Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1).