Deep learning as a tool for early cinema analysis

Bhargav, S.; van Noord, N.; Kamps, J.

DOI
10.1145/3347317.3357240

Publication date
2019

Document Version
Author accepted manuscript

Published in
SUMAC '19

Deep Learning as a Tool for Early Cinema Analysis

Samarth Bhargav
samarth.bhargav92@gmail.com
University of Amsterdam
Amsterdam, The Netherlands

Nanne van Noord
n.j.e.vannoord@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Jaap Kamps
kamps@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

Visual Cultural Heritage has extensively been explored using multimedia methods, but has so far been limited to still images. In particular, Early Cinema has hardly been explored. We analyze the *Desmet* collection, a recently digitized collection of early cinema (1907-1916), in the context of *intertitles*. Intertitles played an important role in silent movies in order to convey the main narratives, and split the film into semantically meaningful segments. We first build several classifiers to detect these intertitles, and evaluate it on a gold standard collection annotated by an expert. We illustrate the usefulness of using Deep Learning methods to extract semantic features to analyze the role of intertitles in early cinema. Furthermore, we attempt to structure and map the narrative progression of a film with respect to the locations at which shots were filmed.

CCS CONCEPTS

• **Computing methodologies** → *Video segmentation*; Visual content-based indexing and retrieval; Cluster analysis.

KEYWORDS

Deep learning; Early cinema; Visual cultural heritage

ACM Reference Format:
Samarth Bhargav, Nanne van Noord, and Jaap Kamps. 2019. Deep Learning as a Tool for Early Cinema Analysis. In *1st Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC '19), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3347317.3357240

1 INTRODUCTION

Large-scale digitization has given scholars access to thousands of hours of film material offering many new research opportunities. Yet much of this potential remains unexplored due to the laborious task of manually annotating film. Luckily, automatic annotation techniques have shown promising results in a wide variety of film analysis tasks on modern material.

Digital cultural heritage has been studied from many perspectives, but so far, this is limited to still shots [21]. Early cinema and silent films have been hardly explored with these techniques.

In this work, we attempt to bridge this gap by focusing on a distinguishing element of early, silent films, Intertitles. Intertitles have been called "key viseotextual elements of the silent screen" [11]. They are an integral part of the narrative, and provide an interesting way of looking at construction of the narrative in early cinema. Intertitles also play a very important role: they are a visual marker of semantics, for example adding context to scenes through textual narration, or marking transitions between scenes or physical locations. Intertitles come in various forms, with various backgrounds, and can be presented in the form of letters or telegrams. In this work we explore various classifiers to detect intertitles, using both 'Conventional' Computer Vision methods, and Deep Learning models.

While 'conventional' Computer Vision methods work well for detecting most basic intertitles, it requires a large amount of hand-engineering to deal with the large number of unusual variations. Deep Learning on the other hand is 'end-to-end', and requires only data to learn from. In addition, feature extraction methods are limited when it comes to *semantic* analysis, even if they perform well for detecting intertitles.

We make the two key contributions: First, we build several classifiers for detecting intertitles, and list the advantages and disadvantages associated with each of them. We show that Deep Learning has better performance and Deep Learning models allows us to perform semantic analysis of the films. Second, we use Deep learning models to analyze the frames surrounding detected intertitles to uncover (a) the types of intertitles and (b) the narrative structure of the intertitle with respect to the locations used in a silent film.

The paper is organized as follows: In Section 2 we given an overview of related work dealing with the analysis of film and cultural heritage material. Section 3 describes the construction of the datasets we use, the baseline and Deep Learning methods for detecting intertitles; Section 4 explains the experimental setup and evaluation metrics; Section 5 contains the results of our experiments; Section 6 discusses these results; Finally, 7 concludes the paper and outlines future work.

2 RELATED WORK

In the following we will briefly discuss work concerning the visual analysis of cultural heritage material, and the challenges this poses as compared to the analysis of contemporary material. Subsequently, we will more generally discuss work concerning automated film analysis, and how our work relates to these works.

2.1 Visual Cultural Heritage Analysis

Accelerated by the large-scale digitization efforts of cultural institutions, increasing attention is given to the development of multimedia methods for cultural heritage material [8, 16–18, 20, 21]. However, the majority of these works consider the analysis of still images such as paintings and drawings [18, 20], scanned newspaper pages [21], and to a lesser extent comic books [22]. Nonetheless,

SUMAC '19, October 21, 2019, Nice, France

ACM ISBN 978-1-4503-6910-7/19/10…$15.00
https://doi.org/10.1145/3347317.3357240

these relate to ours in that they explore how and to what extent approaches developed and trained on contemporary material can be re-purposed for historical material or material which is visually distinct from typical training data. Across various datasets, data types, and previous works, the potential of building on top of a pre-trained deep learning model has been shown [16–18], which informs our choice for how to develop and train our model when applied to silent film material.

2.2 Automated Film Analysis

Analyzing film material from a multimedia perspective has a relatively long and rich history, with numerous works dedicated to the analysis of video on a coarse and fine-grained level [4, 9, 23]. Common tasks include video classification, action and scene recognition, and scene and shot boundary detection. Our work relates most strongly to work on scene and shot boundary detection. Specifically, shots in video are considered elementary units of continuous frames which are typically homogeneous in terms of semantics and appearance [5]. Scenes, on the other hand usually consist of multiple shots typically in the same location and concern a part of the narrative [15]. In literature the automatic detection of shots has received more attention, due to the possibility of detecting shot boundaries based on changes in visual appearance, whereas scene boundaries require analysis of the narrative. However, while shot boundary detection is certainly useful, reliably being able to detect scene boundaries would be an incredible technical feat, which would create tremendous potential for semantic and narrative analysis of video and film material.

Intertitles are added to films in editing, and may occur at scene and shot boundaries, but might also occur at any other point during a shot or scene. Nonetheless, they share two properties with shots and scenes which we exploit in this work, firstly, intertitles are visually distinct from the actual recorded material making them easier to detect than scene boundaries, and secondly, intertitles have an important narrative role, making them more meaningful than shot boundaries. To this end, in this work we develop an approach for detecting intertitles, and subsequently use these detected locations in the film to analyze the content surrounding the intertitle to explore whether it is possible to narrow down the narrative role of an intertitle.

3 MATERIAL AND METHODS

Any analysis of intertitle usage in films requires first the detection of intertitles. We therefore need a classifier which classifies each frame of a film into two categories: `intertitle` or `not intertitle`. Any classifier, whether it uses 'conventional' Computer Vision or Deep Learning, benefits from annotated data, from which it is able to learn a statistical model (or fine-tune hyperparameters). We first gather training data, with frames labeled either as `intertitle` or `not intertitle`. This collection is aided by detecting shots and picking the first few frames, as described in Section 3.2. We first test our method on two baselines (Section 3.3), which use conventional Computer Vision approaches. We outline their limitations and implement a Deep Learning (Section 3.4) algorithm that overcomes some of these limitations. We further construct a 'Gold Standard' collection which consists of 25 films, annotated by experts. This allows us to test the classifiers on 'real' data, as opposed to the balanced annotation dataset.

3.1 The Desmet Collection

The Jean Desmet collection is an archive of around 900 films produced between 1907 and 1916, in addition to several documents, posters and photos. This collection has now been digitized and efforts to explore it using data driven techniques have been carried out [12]. This collection has is key importance to research on early cinema, partly because it has several films from the transitional phase of early cinema. The EYE Film museum has released a part of the collection on YouTube, which we use in our experiments. Of these, we use 25 movies as a 'Gold Standard' collection (see 3.2), and the rest (206 films) to collect annotation data used to train the frame level classifiers (Section 3.2).

3.2 Annotation Data

To facilitate the annotation process, we use `py-scene-detect` [1] to detect shot boundaries in the videos. As intertitles consist of interruptions in the stream of recorded frames with a strongly differing visual appearance, the shot boundary detection algorithm detects them as shot boundaries. By exploiting this property of the shot boundary detection algorithm we are able to bootstrap the annotation process, and perform annotation on a shot level, rather than a frame level, which reduces the annotation time significantly. The annotation process resulted in a total of 2539 segments, of which 239 are intertitles. We use a 60%/20%/20% split into **train**/**val**/**test** set, respectively.

Note that the shots detected by the algorithm are *noisy* in that there are several shot boundaries even within one shot. This is perhaps because of the noise present in the digitized films, and perhaps because using this method requires extensive tuning of the parameters for high efficacy. While this reduced the effort of annotating the data, we therefore avoided using this as a benchmark.

Gold Standard Collection. Each of the 25 'test' films were hand-annotated by an expert. This was done primarily on the ELAN annotation tool [2], which outputs a `.EAF` file per film.

3.3 Baselines

We use two baselines to test the annotation algorithm. Note that the best model among these are selected based on the annotation data (Section 3.2), and this model is tested on the gold standard data (Section 3.2).

3.3.1 Color Histogram Baseline. Color histograms are a simple way to capture the color distribution in an image. Since intertitles are black with white text in the foreground, color histograms should be discriminative for the intertitle classification task. However, since early films are noisy (in the stock itself or acquired during the digitization process), and some of the movies have dark backgrounds, we can expect a high number of false positives. Color histograms have the advantage of being simple and extremely fast to compute.

3.3.2 Text Detection. The presence of text in a frame can indicate an intertitle, since intertitles usually have some text in the image.

[1] http://py.scenedetect.com/

However, since we are looking at early cinema, some of the text are highly stylized, and might be hard to detect using text detection algorithms trained on text.

3.4 Deep Learning methods

The Color Histogram and Text Detection methods are relatively simple and interpretable, but they suffer from drawbacks: They require 'feature engineering', or modifying the algorithm in some way to suit only the task at hand. For instance, for the Desmet collection that we acquired from YouTube, we had to filter detected text from certain regions of the image (see Section 4.2 for more details). They also preclude analysis of the *content* of the image i.e semantic analysis - since they are computed on the pixels of the image and not on the content.

We therefore use Deep Learning to learn to detect intertitles in an *end-to-end* manner. In addition, we explore semantic analysis of the content of the film itself. Deep Learning, in particular Supervised Deep Learning, involves learning a *neural network* from observations of (x, y), where x is the data and y is the target we want to learn. In our case, x is a frame we want to classify, and y is a binary variable indicating whether the frame is an intertitle. Having learned a model from a collection of (x, y) pairs, we can then use this model to predict the target variable y' for a data point we haven't seen before (x'). Deep Learning methods have enjoyed great success in Computer Vision in classifying images [7] and have also been successfully applied to Digital Humanities [17, 20, 21].

As Deep Learning typically requires several thousand data points to learn a model, we resort to Transfer Learning [13], a technique that allows us to 'fine-tune' an existing model to a different task in a similar domain. This involves taking a model that has already been trained on a large dataset and fine tuning the network to perform well on a smaller (but similar) dataset.

4 EXPERIMENTAL SETUP

4.1 Color Histogram Baseline

We use OpenCV [1] to compute a $16 \times 16 \times 16$ color histogram of each frame, resulting in a 4096 dimensional feature vector. We then use a Logistic regression classifier from the `scikit-learn` library [14] that uses the histogram features to predict the presence of an intertitle. We achieve an accuracy of 92.93% on the test set of the annotation data.

4.2 Text Detection Baseline

We use a pre-trained EAST [25] model, implemented in OpenCV. For each frame, we first resize the image to the pre-trained model's required size of 320×320, and run the text detection algorithm. The algorithm outputs multiple bounding boxes, and a confidence score associated with each box. We pick the confidence threshold by picking the threshold which scores the best on the train set. Since the frames we gather have watermark with text in them (see Figure 1b), we exclude boxes detected in the the top right and bottom left regions in the image. Then we classify a given frame as an 'intertitle' if there is at least one box in the image. The final model achieves an accuracy of 89.03% on the test set of the annotation frames.

4.3 Deep Learning methods

For our transfer learning setup, we use a Inception V-3 model [19] trained on the ImageNet dataset [3]. We remove the output classification layer, and replace it with a Softmax layer of 2 classes. In addition, we freeze all layers before the `Mixed_7a` layers. We use an Adam optimizer [6] and train the model with a batch size of 32, for 10 epochs. The model achieves an accuracy of 99.80% on the test set of the annotation frames.

4.4 Prediction and Smoothing

The output of the frame level classifiers are a probability per frame - which by itself is a very noisy signal. We therefore apply a smoothing operator which computes a moving average over n (we use $n = 5$) frames. Note that this method is applied to all methods, including the baselines.

4.5 Evaluation metrics

The evaluation metrics are computed by comparing the output of each algorithm on the Gold Standard set. There is a trade-off between precision (which is the fraction of correctly detected instances over all detected instances), and recall (which is the fraction of detected intertitles over all intertitles encountered). If a classifier has high precision, it implies that each frame that it detects as an intertitle is likely to be an actual intertitle (at the cost of not detecting some frames as an intertitle); while a high recall classifier detects most intertitles, at the cost of some false positives.

In this work, we focus on **recall** i.e., we want to detect all or most of the intertitles, at the cost of some non-intertitle frames being detected as an intertitle. Note that a change to a high precision model (or high F1) is trivial to make.

4.6 Semantic Analysis

The use of Deep Learning models enables us to perform semantic analysis of the content of the film. In particular, we focus on the fact that intertitles play an important semantic role - they usually are placed at strategic positions in film, highlighting a scene change or having an expository function.

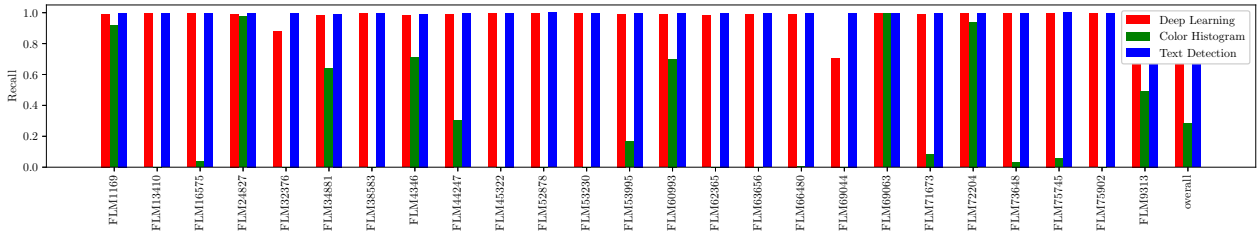
4.6.1 Exploring the role of intertitles. In this analysis, we explore the role of an intertitle in the context of its function, by looking at the frames just before and after an intertitle. If there is a great difference in the content of the scene, then we can categorize this as a intertitle which introduces the next scene. Otherwise, if the content remains more or less the same, then the intertitle perhaps plays an expository role, or contains dialogue. We compute features from the frames before and after the intertitle and compute the difference between them. This difference vector is indicative of the difference in 'content' of the frames: if high, it points to the content of the image being different; if low, it means that the difference is not great. To visualize this, we reduce the dimensions of this vector to 2D dimensions using t-SNE [10].

Since the network in Section 4.3 is tuned for detecting intertitles, and may not capture the semantic content we want to capture in this experiment, we use a Inception-V3 network trained on the ImageNet dataset instead. In particular, we use the features from

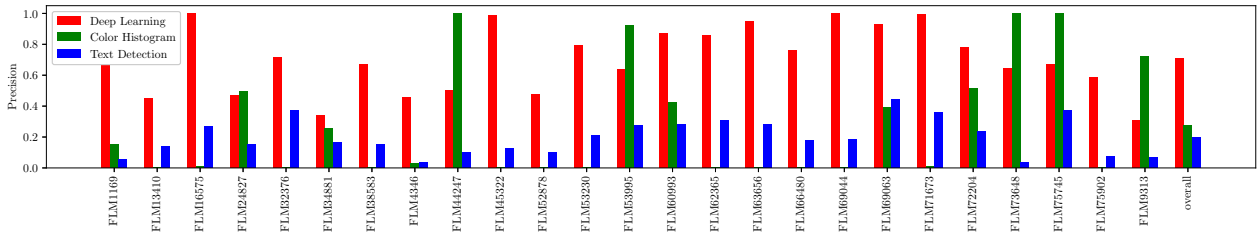
(a) An example intertitle

(b) Detected text in the intertitle frame. Note that we filter out any recognized text in the top right and bottom left areas of the image (which contain logos).

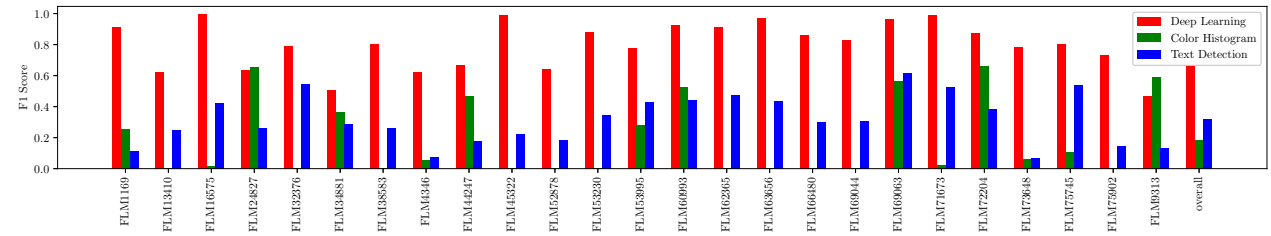
Figure 1: An example intertitle (left), and a visualisation of the automatically detected text (right).



(a) Recall



(b) Precision



(c) F1 Score

Figure 2: Results of all 3 methods. Note that both the Deep Learning method and text detection method perform well in terms of recall, but the text detection method has much poorer precision

Table 1: Average Precision, Recall and F1 scores for the 3 methods. The best performance for a given metric is in bold

Method	Recall	Precision	F1
Color Histogram	28.23	27.7	18.51
Text Detection	**99.64**	19.91	31.71
Proposed Method	97.57	**70.75**	**79.84**

just before the fully connected layers. This yields us feature vectors $\in \mathbb{R}^{2048}$

4.6.2 Analysis using PlacesCNN. In addition, we use a model trained on the Places365 [24] dataset to recognize scenes in the the frames that occur just before and after an intertitle. This method provides us an opportunity to explore which unique scenes are present in the movie - are they 'visited' again in the future? Or is there a 'linear' in the context of the scenes in the movie? We visualize these scenes by projecting the features of these detected scenes into 2D space and drawing arcs from 'before' and 'after' frames. This allows us to uncover the narrative structure of the film.

5 RESULTS

5.1 Intertitle Detection

The results of the two baselines and the deep learning method is reported in Table 1. The precision, recall and F1 score for each of the 25 movies (and an average) can be seen in Figure 2. The color histogram has poor performance compared to the other methods. The text detection method has the best recall, but has the lowest precision among the 3 models, which makes this unusable in practice. The Deep Learning model has comparable recall with the text detection method, but has a much higher precision and overall better F1 score. We therefore resort to use the Deep Learning method to detect intertitles which are then used to perform semantic analysis.

5.2 Semantic Analysis

5.2.1 Exploring the role of intertitles. Figure 3 is a t-SNE plot of the difference vectors that we computed on the entire Desmet collection, including samples from some regions in the clusters. There seems to be 3 distinct clusters. We observed the cluster on the left to contain the 'start' intertitle which has the title of the film. The in-frame just before these detected intertitle is almost black and the out-frame has some content, resulting in a large difference. Similarly the rightmost cluster consists of the 'end' shots of movies. The cluster in the middle is diverging, and we found a lot of them to be the result of noisy detection - usually letters or missives which aren't always detected as intertitles, which have the same in/out frames. This is to be expected since we used a model optimized for for high recall. As we move outward, the difference in the in/out frames become pronounced.

5.2.2 Analysis using PlacesCNN. Figure 4 contains the analysis for a single film. We manually identified the 3 unique locations in this film, which we label 'Indoors (1)', 'Cafe' and 'Indoors (2)'. The t-SNE plot of the features shows that the scenes are clustered together according to the location. For instance, the orange dots are clustered together in top top of the figure, and correspond to 'Indoors (1)'. The one 'Indoors (1)' image that occurs on the bottom left has several people occluding the scene, which leads it to be placed farther away. Note also that the location 'Indoors (2)' is placed far away from the other points. Although we illustrated one film by manually annotating the unique locations in the film, analysis of other movies was done fully 'unsupervised'.

6 DISCUSSION

6.1 Detecting Intertitles

The Color Histogram baseline performs very poorly on the gold standard collection, despite having a high performance on the annotation data. This shows that it overfits on the annotation data, even though we use a very simple Logistic Regression model. The text detection model has nearly perfect recall, but this comes at a cost: it has a very low precision of around 20%. This means that only 1 out of 5 frames that it classifies as an intertitle is actually an intertitle i.e it has a lot of false positives (see Figure 5).

The Deep Learning method however, has a high precision (even though the best models were selected based on recall) and overall the highest F1 score. Perhaps the only disadvantage of applying Deep Learning is the time taken during inference: it is indeed much slower than the color histogram method. However, this can be mostly be offset with the use of GPUs.

The fine-grained performance for the 3 algorithms we use for detecting intertitles can be seen in Figure 2. Note that the the Deep Learning model performs consistently for all films: The Color Histogram seems to work well only for some films. The text detection model has very low precision for almost all films, indicating that it just predicts several frames as intertitles

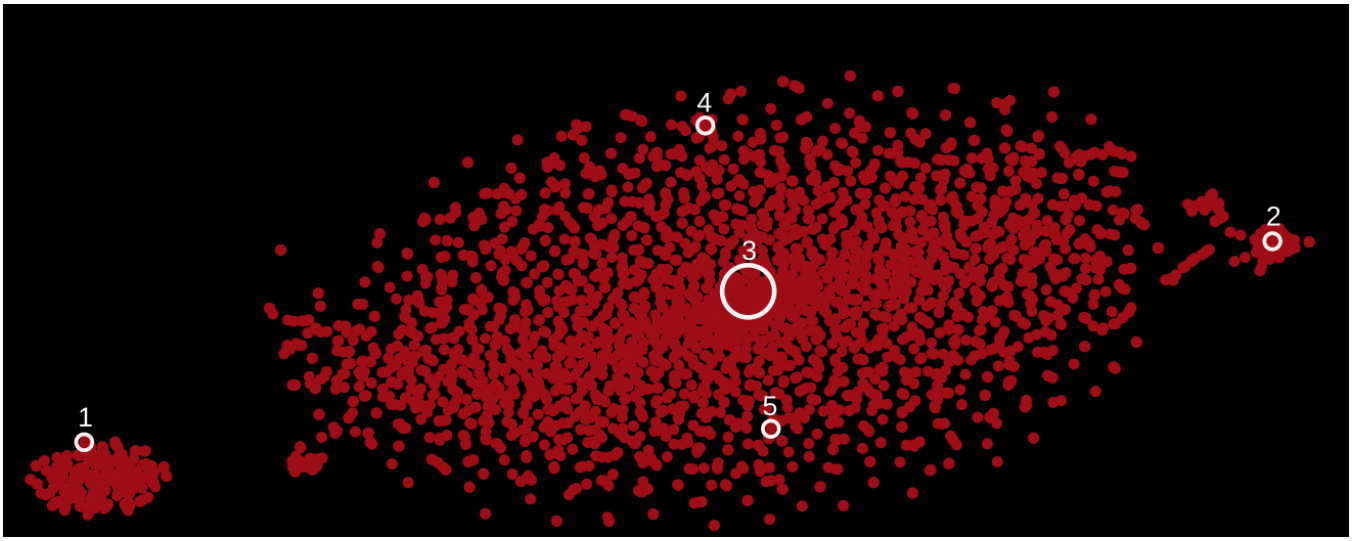
6.2 Semantic Analysis

In the t-SNE plot of the difference vectors (Figure 3), we see a clear pattern emerge. The intertitles are grouped into clear clusters, and this is done in an 'unsupervised' manner. This is possible only with methods which extract 'high-level' features such as the presence of objects (which is what the model we used, Inception-V3, is trained to do), instead of 'low' level features like Color Histograms. We note that this is not possible at all if we use the text detection method: there is no way to extract semantically useful features.

The second use of a Deep Learning model involves applying a Deep Learning model trained on the Places365 dataset, which classifies a given image into scene categories. This allows us to uncover the narrative structure of a film in the context of the location. In Figure 4d, we can see that the model indeed captures the 3 locations in the movie. We note that like the previous method, this is also unsupervised, and this kind of analysis cannot be performed without semantic information about the locations.

7 CONCLUSION

In this work we have shown the advantage of using Deep Learning models for the detection and analysis of intertitles. Deep Learning models generalize very well and have superior performance; they are data efficient when transfer learning is employed; they give us access to high level semantic features otherwise not possible using conventional Computer Vision methods. By presenting a reliable



(a) t-SNE plot of the difference (of feature vectors) of the before/after frames

(1) An example of a intertitle that signals the start of the film

(2) An example of a intertitle that signals the end of the film

(3a) Sample from (3)

(3b) Sample from (3)

(4) Sample from (4)

(5) Sample from (5)

Figure 3: t-SNE plot of the difference (of feature vectors from a pre-trained Inception-V3 model) of before/after frames. Notice the two distinct clusters on the extreme left and right - they are the start (left) and end (right) intertitles that our model detected

and accurate approach for the detection of intertitles we open the door for future and richer automated analysis of early cinema film material and film narrative in general.

ACKNOWLEDGMENTS

The moving image data is publicly available from the EYE Film museum. All code and frame-level annotations, including the novel benchmark for intertitle detection, are available from https://github.com/samarthbhargav/DRAFT. The authors would like to thank Rob Wegter for his input on Early Cinema throughout the project and for annotating the Gold Standard collection.

REFERENCES

[1] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[2] Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating Multimedia/Multi-modal Resources with ELAN.. In *LREC*.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. 1995. Automatic Recognition of Film Genres. In *Proceedings of the Third ACM International Conference on Multimedia (MULTIMEDIA '95)*. ACM, 295–304.

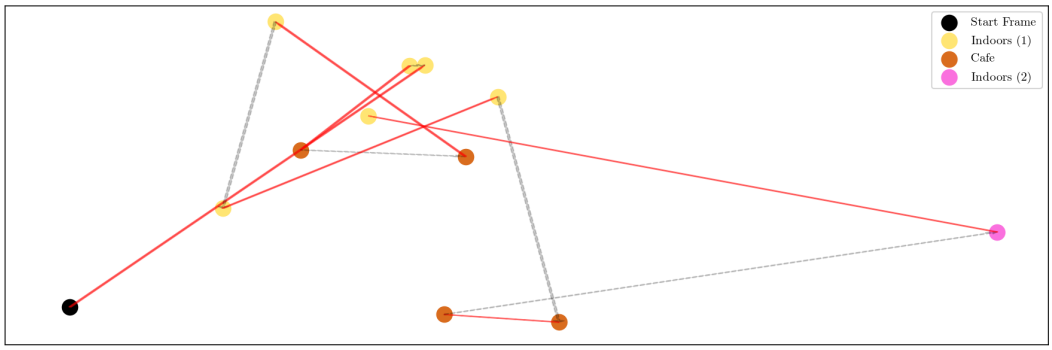
[5] A. Hanjalic. 2002. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology* 12, 2 (Feb 2002), 90–105.

[6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[8] Jia Li, Lei Yao, Ella Hendriks, and James Z Wang. 2012. Rhythmic brushstrokes distinguish van Gogh from his contemporaries: findings via automated brushstroke extraction. *IEEE transactions on pattern analysis and machine intelligence* 34, 6 (jun 2012), 1159–1176.

(a) Indoors (1) (b) Cafe (c) Indoors (2)



(d) t-SNE plot of the features of all frames before/after an intertitle occurs. The red lines connect frames separated with intertitles and the black lines connect the first frame and last frame in a segment. Notice how visually similar scenes are clustered closer to each other. The left most yellow point (Indoors (1)) is farther away from the other yellow points, perhaps because of the scene being occluded by several people

Figure 4: The different locations in the film 'De Drankduivel' (lit. 'The Drinking Devil')

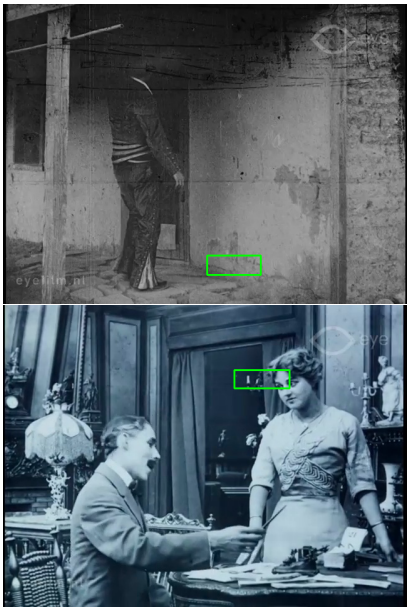


Figure 5: False Positives for the text detection method

[9] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. ACM, 843–851.

[10] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[11] Katherine Nagels. 2012. Those funny subtitles: Silent film intertitles in exhibition and discourse. *Early Popular Visual Culture* 10, 4 (2012), 367–382.

[12] Christian Gosvig Olesen, Eef Masson, Jasmijn Van Gorp, Giovanna Fossati, and Julia Noordegraaf. 2016. Data-driven research for film history: exploring the Jean Desmet collection. *Moving Image: The Journal of the Association of Moving Image Archivists* 16, 1 (2016), 82–105.

[13] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[15] Z. Rasheed and M. Shah. 2003. Scene detection in Hollywood movies and TV shows. In *Conference on Computer Vision and Pattern Recognition.*, Vol. 2. II–343.

[16] Babak Saleh and Ahmed Elgammal. 2015. Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. *arXiv* (2015), 21. arXiv:1505.00855

[17] Gjorgji Strezoski, Nanne van Noord, and Marcel Worring. 2019. Learning Task Relatedness in Multi-Task Learning for Images in Context. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR '19)*. ACM, 78–86.

[18] Gjorgji Strezoski and Marcel Worring. 2018. OmniArt: A Large-scale Artistic Benchmark. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 4, Article 88 (2018), 21 pages.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015). arXiv:1512.00567

[20] Nanne van Noord, Ella Hendriks, and Eric Postma. 2015. Toward Discovery of the Artist's Style: Learning to recognize artists by their artworks. *IEEE Signal Processing Magazine* 32, 4 (2015), 46–54.

[21] Melvin Wevers and Thomas Smits. 2019. The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities* (2019).

[22] Kim Young-Min. 2019. Feature visualization in comic artist classification using deep neural networks. *Journal of Big Data* 6, 1 (25 Jun 2019), 56.

[23] Zhaoyang Zhang, Zhanghui Kuang, Ping Luo, Litong Feng, and Wei Zhang. 2018. Temporal Sequence Distillation: Towards Few-Frame Action Recognition in Videos. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. ACM, 257–264.

[24] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[25] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. EAST: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5551–5560.