Computational segmentation of Wayang Kulit video recordings using a Cross-Attention Temporal Model*

Hong Wei Shawn Liew[1], Miguel Escobar Varela[2,3]

[1]Faculty of Science, National University of Singapore
[2]Faculty of Arts and Social Sciences, National University of Singapore
[3]Centre for Computational Social Science and Humanities, National University of Singapore

Abstract

We report preliminary findings on a novel approach to automatically segment Javanese wayang kulit (traditional leather puppet) performances using computational methods. We focus on identifying comic interludes, which have been the subject of scholarly debate regarding their increasing duration. Our study employs action segmentation techniques from a Cross-Attention Temporal Model, adapting methods from computer vision to the unique challenges of wayang kulit videos. We manually labelled 100 video recordings of performances to create a dataset for training and testing our model. These videos, which are typically 7 hours long, were sampled from our comprehensive dataset of 12,638 videos uploaded to a video platform between 03 Jun 2012 and 30 Dec 2023. The resulting algorithm achieves an accuracy of 89.06% in distinguishing between comic interludes and regular performance segments, with F1-scores of 96.53%, 95.91%, and 92.47% at overlapping thresholds of 10%, 25%, and 50% respectively. This work demonstrates the potential of computational approaches in analyzing traditional performing arts and other video material, offering new tools for quantitative studies of audiovisual cultural phenomena, and provides a foundation for future empirical research on the evolution of wayang kulit performances.

Keywords

video processing, temporal models, performing arts, wayang kulit

1. Introduction

Scholars in the digital humanities have increasingly turned their attention to the automated examination of video materials [1, 15, 4, 9, 16, 11]. However, the majority of these efforts have been concentrated on North American and European content, primarily focusing on film and television. One particularly promising area for expansion is the study of other cultural phenomena from around the world, such as recorded theatrical performances. This paper aims to contribute to this expansion by presenting a novel approach to analyzing video recordings of Javanese wayang kulit (shadow puppet theater) performances using computational video segmentation techniques.

Javanese wayang kulit, a centuries-old theatrical tradition, adheres to a meticulously structured sequence of scenes that forms the bedrock of every artist's training and practice. This rigorously codified structure, which has been extensively studied and documented [2, 5, 7], includes as essential components two comic interludes: limbukan and gara-gara. These segments,

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
shawnliew@u.nus.edu (H. W. S. Liew); m.escobar@nus.edu.sg (M. E. Varela)
0009-0000-2097-9074 (H. W. S. Liew); 0000-0001-8396-1664 (M. E. Varela)

characterized by their improvisational nature, contemporary references, and often irreverent humor, serve as a counterpoint to the more formal, plot-driven portions of the performance while remaining integral to its overall structure. Typically lasting at least 30 minutes each within the approximately 7 hour duration of a full performance, these interludes have become a subject of debate among practitioners and scholars alike. While there is consensus on the precise moments these interludes begin and end, opinions diverge on whether their perceived increasing length is a positive or negative development [17, 7]. This debate rests on an underlying assumption shared by both critics and supporters: that the comic interludes are indeed becoming longer. Given the vast number of performances that occur annually (see below) and their considerable duration, empirically verifying this claim poses a significant challenge.

This paper aims to address this gap by applying computational analysis to a corpus of wayang kulit recordings, offering a data-driven perspective on this longstanding debate. We applied a state-of-the-art cross-attention temporal model and approached our problem as an action segmentation task to take advantage of cutting-edge transformer architectures for automated video analysis. By training such a model on a culturally-specific dataset, we seek to create a tool that can accurately detect these segments across a large corpus of performances available on platforms like YouTube. This approach not only allows us to quantitatively assess the changing duration of comic interludes over time but would also enable us to explore potential correlations with other factors, such as the performance location, the profile of the dalang (puppeteer), or the video's popularity as measured by view counts. In this pilot study, we demonstrate the potential of computational video analysis for studying traditional performing arts and contributing to ongoing debates that matter to the scholars and practitioners of these traditions. Furthermore, this research serves as a case study for the broader application of computational methods to video content in the humanities. By addressing the unique challenges posed by culturally specific, long-form performances, we hope to pave the way for similar analyses across a wide range of performance traditions and visual media.

In our initial exploration, we manually labelled 100 videos using a custom interface. These videos, which are typically 7 hours long, were sampled from our comprehensive dataset of 12,638 videos (see below for details). By training the Frame-Action Cross-Attention Temporal Modeling for Efficient Action Segmentation (FACT) model [14] on our dataset, we obtained an accuracy of 89.06% and an F1 score of 92.47% with an overlap threshold of 50%. These results are particularly remarkable given the relatively small dataset used for training and the cultural-specificity of the content. These encouraging results show the enormous promise that similar models have for the analysis of audiovisual media using computational methods.

In the remainder of this paper, we will outline our methodology, our current results, and future directions. An illustration of our workflow is presented in Figure 1.

2. Dataset

2.1. Assembling the Dataset

As a first step to identify and quantify the duration of the comic interludes, we assembled a dataset of YouTube videos. We used the YouTube Data API to obtain all videos on YouTube with the keyword "wayang kulit". We conducted this search on 30 December 2023 and created a list

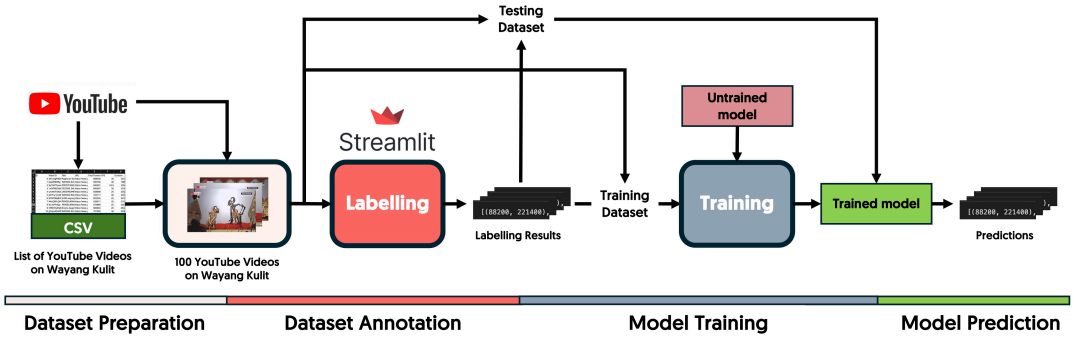

Figure 1: Illustration of entire workflow.

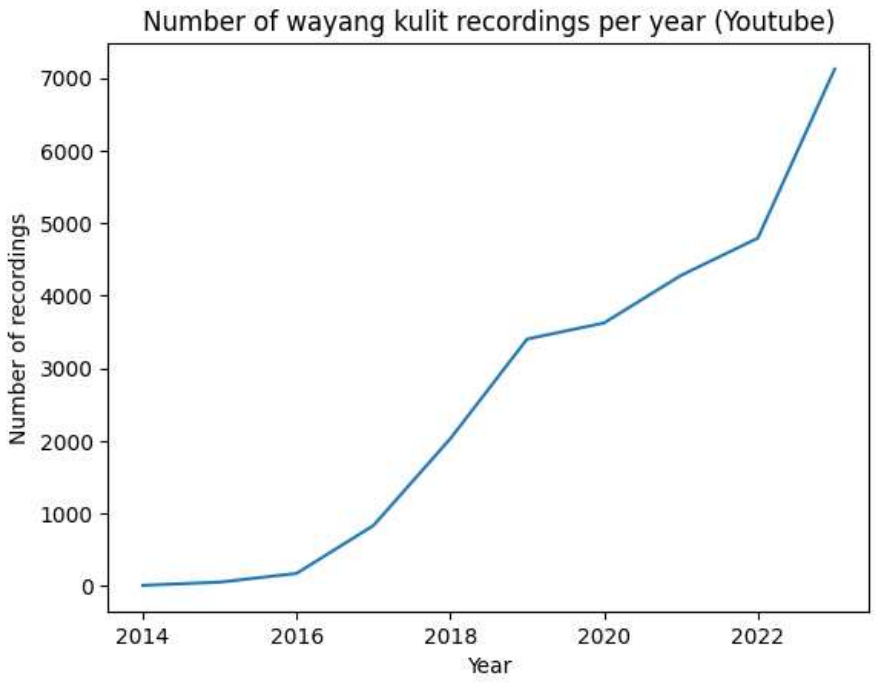

Figure 2: Number of wayang kulit recordings per year.

of all existing videos. The total count was 34,349 items. We found that the earliest videos that could be retrieved from the API were from 2012. We used the descriptions and dates to identify potential duplicates but found none. We also excluded videos that were too short (less than 5 hours), which would represent excerpts of the performances, rather than full performances, or versions of wayang kulit from Bali (Indonesia) or Kelantan (Malaysia) which are shorter than Javanese wayang kulit. After these steps, we had in our hands a comprehensive dataset of 12,638 videos that were uploaded between 03 Jun 2012 (inclusive) and 30 Dec 2023 (inclusive).

For each video, we also retrieved the following metadata:

- Title
- Video ID
- Channel Name
- Published Date
- Duration
- Description
- Number of Views
- Number of Likes
- Number of Comments

As Figure 2 shows, the number of videos has increased over time. This is probably due to the increased availability of high-quality video-cameras in consumer phones and the decreasing costs of internet access in Indonesia. However, it does not necessarily mean that the number of performances has increased. Evidence from social media collected earlier [8, p. 156] suggest that the number of performances for the past few years is steady (about 3700 per year).

For this initial experiment, we constructed a sample of 100 recordings for labelling and analysis. To ensure our sample approximates the temporal distribution of actual performances, and not just performances captured in our dataset, we employed a stratified sampling technique based on year. This method allowed us to maintain the proportional representation of performances from different years, mitigating the risk of oversampling more recent videos simply due to their greater availability online. This approach not only helps us develop a robust model for segmenting wayang kulit videos but also ensures that our initial findings are based on a representative sample of performances across the decade under study. The results from this preliminary analysis will inform our subsequent application of the model to the broader dataset, enabling us to draw more comprehensive conclusions about the evolution of comic interludes in Javanese wayang kulit performances.

2.2. Labelling the Dataset

A wayang spectator can readily recognize comic interludes when viewing a sequence of frames. As noted earlier, a typical wayang kulit show includes two comic interludes: the first is called *limbukan* and the second is called *gara-gara*. These are long interludes, and typically last at least 30 minutes each (although the interludes are perhaps getting longer, as we eventually aim to determine). These comic interludes in wayang kulit typically feature the following elements:

1. Specific puppets: Limbuk and Cangik for *limbukan*; Petruk, Bagong, Gareng, and sometimes Semar for *gara-gara*.
2. Interactions between the performers and the audience.
3. Guest performers on stage.
4. Extended solo performances by the *pesindhen* or female singers.

A full description of the elements in these comic interludes can be found in [7]. It's important to note that these elements in isolation are not definitive indicators of a comic interlude. For instance, characters like Petruk may appear in other scenes, and female singers perform

throughout the show. However, during comic interludes, these elements are more seen in combination. Guest performers usually signify a comic interlude, but they may also appear before the show begins, adding another layer of complexity to the identification process. These nuances underscore the importance of considering the sequence of frames rather than relying on isolated visual cues or single-frame classification approaches. A simple detection algorithm to identify specific characters (e.g., Petruk) would be insufficient, especially given the regional and artistic variations in puppet design. While creating a comprehensive dataset of these variations could be valuable, it falls outside the scope of the current project.

To address these challenges and leverage culturally-specific knowledge, we adopted the following approach for labelling our data:

1. Thumbnail Generation: We extracted thumbnails at 15-second intervals from each of the 100 selected videos in our dataset. The interludes under consideration are usually at least 30 minutes long, so this sampling rate is justified.
2. Visual Interface: We developed a custom interface using Python and the Streamlit library [18] to display these thumbnails on a grid. This layout allowed for efficient visual scanning of the performance timeline.
3. Expert Labeling: An expert in wayang kulit (who is one of the present authors) manually reviewed the thumbnail grids for each video. The expert identified and marked the frames that fell within comic interludes, considering the context provided by surrounding frames.
4. Sequence Preservation: By using a grid display of sequential thumbnails, we enabled the expert to consider the progression of the performance, crucial for accurate identification of comic interludes.
5. Start and End Points: The expert marked the beginning and end of each comic interlude, allowing us to capture the duration and placement of these segments within the overall performance.

This methodology allowed us to create a labelled dataset that captures the contextual cues that mark the comic interludes. By preserving the sequential nature of the performances in our labeling process, we aim to develop computational models that can more accurately detect these segments. Screenshots of the interface developed for this purpose are shown in Figure 3.

Having a single person create the labels might sound problematic, but it should be noted that the markers of a comic interlude are very extremely rigid, as noted above. To give readers not familiar with wayang kulit a crude approximation, imagine a dataset of TV shows that also includes commercial breaks. Imagine you are looking at a series of still images from this dataset. At first glance, it might be hard to tell which images are from the show and which are from commercials. For example, you might see a frame showing a person walking on the street, which could be from either the show or a commercial. However, if you look at the sequence of images, patterns emerge that make it easy to spot where the commercials begin and end. This is similar to our wayang kulit performances. (It's worth noting that this analogy isn't perfect; in reality, ads typically show the products they're advertising in a unique visual style, which would make them easily identifiable even from a single frame but no such cues are available in our context). The key point is that, like identifying commercial breaks in a sequence of TV show thumbnails, someone with even passing familiarity with wayang kulit can easily

Figure 3: Labelling page of our labelling interface. Upon loading the thumbnails, we presented the thumbnails chronologically in a grid. This layout provides the user with an efficient overview of the video.

determine which segments of a performance correspond to comic interludes by looking at the overall structure, even without audio or full video. Readers interested in looking at an actual wayang kulit show with a clearly marked scene structure, can refer to [2]. Its accompanying website includes a full video recording with a comprehensive transcription, translation and description of an entire wayang show.

During the labeling phases, we identified 3 videos that didn't match our expectations: one was just an audio recording with a still image, and the other two were just collections of short performances, rather than a full-length wayang kulit. In the future, we plan to use computational methods to identify these types of videos, but for the time being, we decided to proceed with a dataset of 97 labelled videos.

3. Model training

For the next step in our workflow, we defined our task as follows: Given the labelled footage of wayang kulit shows, our objective is to segment the content into two distinct categories: comic interludes and regular performance. We propose to frame this as a supervised "action segmentation" problem, despite the fact that our categories do not strictly represent actions in the conventional sense. This approach is valuable for several reasons:

1. Temporal continuity: Like actions, these segments unfold over time.

2. Context dependency: Identifying comic interludes often requires understanding the broader context of the performance, similar to how actions are interpreted in a sequence.

The action segmentation problem has been widely studied and some common datasets available for this problem are the Epic-Kitchens [6] and the Breakfast Actions [12] datasets. Conventional approaches to this problem typically employ convolutional-based networks such as MS-TCN [10] and MS-TCN++ [13]. These networks use 1-dimensional dilation convolutions to capture long-range temporal features. However, they are often prone to over-segmentation [14]. On the other hand, two-stage architectures that first calculate frame-based features followed by action features don't take full advantage of the low and high-level features captured by the respective stages [14].

To improve upon the two-stage methods, Lu and Elhamifar (2024) designed a recent state-of-the-art model called FACT (FACT: Frame-Action Cross-Attention Temporal Modeling for Efficient Action Segmentation) [14]. The FACT model improves upon conventional two-stage based approaches by employing two branches in parallel – the action branch and the frame branch – to perform action segmentation. As Lu and Elhamifar (2024) explain, the action branch captures high-level action dependencies with transformers whereas the frame branch captures low-level frame features using dilated convolutional layers. The authors note that since the details captured in both branches are complimentary, they are therefore interconnected using a cross-attention mechanism. The general procedure is as follows:

1. Model initialisation: The frame branch is initialised with the features and is updated once using the dilated convolutional layers. The action branch is then initialised using the above output with action tokens. Action tokens can be thought of as labels for the different segments. In our case, we only have one label that corresponds to the *interlude*. The other frames would belong to the default category.
2. Update sequence:
    a) On the frames branch, the features are passed through the dilated convolutional layers.
    b) On the action branch, a cross-attention mechanism is used to combine the output from the frames branch (generated in Step a) with the action tokens on the action branch.
    c) On the action branch, we used the output (generated in Step b) and advance it through a transformer to capture action dependencies.
    d) On the frames branch, we used another (different) cross-attention mechanism to combine the output from the action branch (generated in Step c) with the features (generated in Step a).
    e) This cycle is then repeated for each block.

An illustration of the FACT model is available in [14] Figure 1. To adapt the FACT model to our context, we carried out the following procedure:

1. We provide the model with thumbnails, that were down-sampled to reduce computational resources needed, sampled at 60-seconds intervals to match the labelling results we obtained in the previous step.

2. With these thumbnails and labelled results, we randomly split our labelled dataset into the training dataset (77 videos) and the testing dataset (20 videos).
3. Once the training dataset was collated, we initialised the training. During the training, we routinely evaluated the performance of our model using the testing dataset. We then selected the model weights corresponding to the highest F1@0.50 score (explained below) on our testing dataset.
4. After obtaining the model weights, we evaluated the dataset using the trained model to obtain our predictions.
5. We repeated the training process across different data splits to obtain reliable metrics.

4. Results

Across the 5 different splits of the training and testing dataset, we obtained an average accuracy of 89.06%. To provide a more nuanced evaluation of our model's performance, we also calculated F1-scores at different thresholds [3]. This is a measure of the similarity between two sequences, in this case, the predicted action segments and the ground truth action segments in a video. Following common practice in the literature, we report the F1 scores (averaged across 5 different splits) at different overlap thresholds: 10%, 25% and 50% as shown in Table 1.

Table 1
F1-values at different temporal tolerances

Overlap Threshold (%)	F1-score
10	96.53%
25	95.91%
50	92.47%

The overlap thresholds refer to the minimum Intersection over Union (IoU) between the predicted and ground truth segments for the prediction to be considered as a True Positive. Selected segments are presented in Figure 4 to illustrate the best three and worst three predicted segments.

5. Future Work

To enhance the accuracy of our wayang kulit segmentation model, we propose several avenues for future research:

1. Expanding our dataset to include a larger number of labelled performances from diverse regions of Java would improve the model's robustness and ability to generalize across different styles.
2. Refining our segmentation approach to differentiate between limbukan and gara-gara interludes, as well as identifying sub-segments within these interludes (song, audience interactions, puppetry segments), could lead to more nuanced and accurate results. Analyzing audience reactions such as laughter and applause could provide additional indicators for detecting comic interludes.

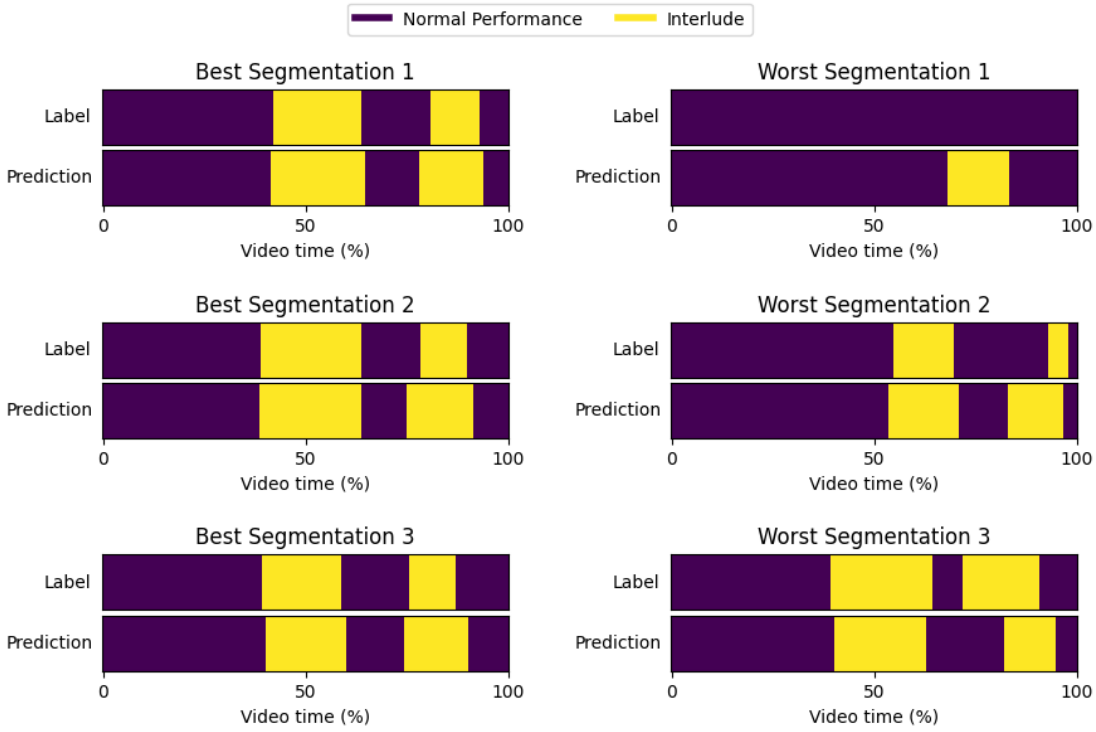

Figure 4: We selected the best three and worst three predicted segmentations (determined visually) by our trained model on our test dataset with binary labels.

3. Integrating multimodal analysis by incorporating audio features to capture musical cues and dialogue would provide additional context for segmentation decisions. A challenge here is that there are no good speech-to-text models openly available for Javanese, the language of the performances.
4. Developing a puppet recognition model to track individual characters throughout the performance could offer valuable visual cues for identifying segment transitions.

An additional challenge for future work lies in precisely locating comic interludes within the overall structure of wayang kulit performances. This task is complicated by the fact that video recordings often include non-performance segments at the beginning or end, such as pre-show preparations or post-performance activities. To accurately determine whether a comic interlude coincides with the end of a performance or begins at a specific point (e.g., 50% through the performance), we must first reliably identify and exclude these non-performance segments. We have conducted preliminary research on this issue using our existing dataset with different labeling schemes. Our initial findings suggest results comparable to our main segmentation task. However, further investigation is needed to develop robust methods for accurately detecting the true start and end points of performances within video recordings and distinguishing between pre-show, post-show, and intermission segments. This line of research is crucial for enabling more nuanced analyses of wayang kulit structure and for tracking potential changes in performance composition over time. As we overcome this challenge and achieve higher

accuracy in our segmentation model, several exciting research directions will become feasible:

1. A longitudinal study could be conducted to track changes in interlude duration over decades, potentially revealing evolving trends in wayang kulit performances. This analysis could be correlated with factors such as performance date, location, and the identity of the dalang (puppeteer) to uncover broader patterns and influences.
2. We could develop a public-facing user-friendly interface that allows wayang scholars and enthusiasts to easily navigate and analyze performances. This tool could revolutionize the study of wayang kulit by providing quick access to relevant segments and facilitating comparative analysis across multiple performances.

Acknowledgments

This work was supported by the Humanities and Social Sciences Seed Fund for Collaborative Research (HSS SFCR) 2023-2024 from the National University of Singapore.

References

[1] T. Arnold and L. Tilton. *Distant viewing: computational exploration of digital images*. Cambridge, Massachusetts: The MIT Press, 2023. URL: https://go.exlibris.link/R2T5GjhB.

[2] B. Arps. *Tall tree, nest of the wind: the Javanese shadow-play Dewa Ruci performed by Ki Anom Soeroto: a study in performance philology*. Singapore: NUS Press, 2016. URL: http://pintubahasa.com/ttnotw/P6-1%5C%5Fvideo.html.

[3] N. Behrmann, S. A. Golestaneh, Z. Kolter, J. Gall, and M. Noroozi. "Unified Fully and Timestamp Supervised Temporal Action Segmentation via Sequence to Sequence Translation". In: *Computer Vision – ECCV 2022*. Ed. by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner. Vol. 13695. Cham: Springer Nature Switzerland, 2022, pp. 52–68. DOI: 10.1007/978-3-031-19833-5_4. URL: https://link.springer.com/10.1007/978-3-031-19833-5%5C%5F4.

[4] M. Burghardt, A. Heftberger, J. Pause, N.-O. Walkowski, and M. Zeppelzauer. "Film and video analysis in the digital humanities–an interdisciplinary dialog". In: *Digital Humanities Quarterly* 14.4 (2020). URL: https://orbilu.uni.lu/handle/10993/45345.

[5] V. M. Clara van Groenendael. *Wayang Theatre in Indonesia: An Annotated Bibliography*. Dordrecht, Holland; Providence, U.S.A: Brill Academic Pub, 1988.

[6] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100". In: *International Journal of Computer Vision (IJCV)* 130 (2022), pp. 33–55. URL: https://doi.org/10.1007/s11263-021-01531-2.

[7] K. Emerson. *Innovation, Style and Spectacle in Wayang: Purbo Asmoro and the Evolution of an Indonesian Performing Art*. Singapore: NUS Press, 2022. URL: https://nuspress.nus.edu.sg/products/innovation-style-and-spectacle-in-wayang.

[8] M. Escobar Varela. *Theater as Data: Computational Journeys into Theater Research*. Ann Arbor: University of Michigan Press, 2021. URL: https://press.umich.edu/Books/T/Theater-as-Data2.

[9] M. Escobar Varela and G. O. F. Parikesit. "A quantitative close analysis of a theatre video recording". In: *Digital Scholarship in the Humanities* 32.2 (2017), pp. 276–283. DOI: 10.1093/llc/fqv069. URL: https://academic.oup.com/dsh/article-abstract/32/2/276/2669634/A-quantitative-close-analysis-of-a-theatre-video.

[10] Y. A. Farha and J. Gall. *MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation*. 2019. arXiv: 1903.01945 [cs.CV]. URL: https://arxiv.org/abs/1903.01945.

[11] B. Flueckiger and G. Halter. "Methods and Advanced Tools for the Analysis of Film Colors in Digital Humanities." In: *DHQ: Digital Humanities Quarterly* 14.4 (2020).

[12] H. Kuehne, J. Gall, and T. Serre. "An end-to-end generative framework for video segmentation and recognition". In: *Proc. IEEE Winter Applications of Computer Vision Conference (WACV 16)*. Lake Placid, 2016.

[13] S. Li, Y. A. Farha, Y. Liu, M.-M. Cheng, and J. Gall. *MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation*. 2020. arXiv: 2006.09220 [cs.CV]. URL: https://arxiv.org/abs/2006.09220.

[14] Z. Lu and E. Elhamifar. "FACT: Frame-Action Cross-Attention Temporal Modeling for Efficient Supervised Action Segmentation". In: *Conference on Computer Vision and Pattern Recognition 2024*. 2024.

[15] L. Manovich. "Visualizing Vertov". In: *Russian Journal of Communication* 5.1 (2013), pp. 44–55. DOI: 10.1080/19409419.2013.775546. URL: http://www.tandfonline.com/doi/abs/10.1080/19409419.2013.775546.

[16] E. Masson, C. G. Olesen, N. van Noord, and G. Fossati. "Exploring Digitised Moving Image Collections: The SEMIA Project, Visual Analysis and the Turn to Abstraction." In: *DHQ: Digital Humanities Quarterly* 4 (2020).

[17] J. Mrázek. *Wayang and Its Doubles: Javanese Puppet Theatre, Television and the Internet*. Singapore: NUS Press, 2019.

[18] Steamlit. *Streamlit: A faster way to build and share data apps*. 2024. URL: https://streamlit.io/.

A. Online Resources

GitHub repository.