Zoetrope – Interactive Feature Exploration in News Videos

Liebl, Bernhard
bernhard.liebl@uni-leipzig.de
Computational Humanities Group, Leipzig University

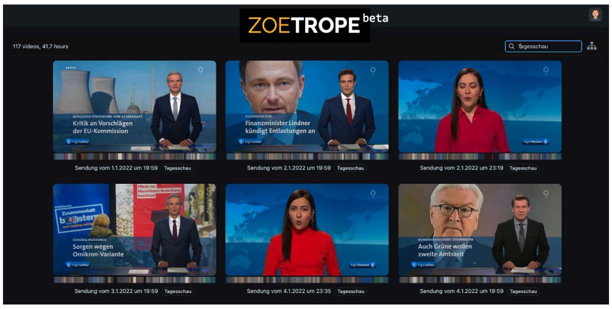
Burghardt, Manuel
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Leipzig University

Introduction

While conceptions such as “Distant Viewing” (Arnold & Tilton, 2019) and “Deep Watching” (Bermeitinger et al., 2019) have recently stressed the importance of developing a methodology for analyzing video material in the digital humanities, there is quite a history of existing tools for the annotation and analysis of videos [1]. At the same time, recent advances in neural networks have paved the way for what has been called the “Visual Digital Turn” (Wevers & Smits, 2020). However, current neural frameworks can be used far beyond the extraction of purely visual features, such as objects (Carion et al., 2020) or faces (Deng et al., 2019). Rather, they are capable of automatically extracting all kinds of multimodal information, including scene text (Chen et al., 2020), speech (Radford et al., 2022), text-image embeddings (Radford et al., 2021) and high-quality automatic image captioning (Li et al., 2022). Oftentimes, these frameworks are cutting-edge technology and require some advanced technical skills to set up. Taking up the basic idea of humanist-computer interaction (Burghardt & Wolff, 2014), which calls for usability engineering and user experience design in the digital humanities, we present the Zoetrope prototype, which is able to process raw features from the above neural frameworks and make them available for interactive exploratory analysis in a web-based tool. Zoetrope currently is built around the use case of analyzing German news videos for narrative strategies (Machill et al., 2007), which eventually will be used to semi-automatically detect patterns of disinformation, so-called *fake narratives*[2] (for some examples see Tseng et al., 2023).

The Zoetrope Prototype

In this section we briefly outline some of the main features of the Zoetrope prototype, which is currently not publicly available. However, to better showcase its current functionality we have prepared a demo video [3]. Zoetrope can be roughly divided in two screens. The first screen provides an overview of all the videos in a collection and allows researchers to select a specific video for more in-depth analyses (see Fig. 1). Videos can be searched and filtered by their tags but also by a keyword search in the OCR (optical character recognition) and ASR (automatic speech detection) tracks, both of which are available for all the videos. The movie barcodes (see Burghardt et al., 2017) underneath each video indicate the most dominant colors along the timeline and can be used to roughly navigate and scrub the videos in a live preview mode.


ZOETROPE beta
117 videos, 41.7 hours
Tagesschau
Kritik an Vorschlägen der EU-Kommission
Sendung vom 1.1.2022 um 19:59
Tagesschau
Finanzminister Lindner kündigt Entlastungen an
Sendung vom 2.1.2022 um 19:59
Tagesschau
Sendung vom 2.1.2022 um 23:19
Tagesschau
Sorgen wegen Omikron-Variante
Sendung vom 3.1.2022 um 19:59
Tagesschau
Sendung vom 4.1.2022 um 23:35
Tagesschau
Auch Grüne wollen zweite Amtszeit
Sendung vom 4.1.2022 um 19:59
Tagesschau

Zoetrope – Overview screen for video retrieval and selection.

Upon clicking a video, researchers are redirected to the interactive feature exploration screen of the selected video (Fig. 2). The screen features a player pane for the selected video and also some controls to navigate it in the navigator and timeline pane. The timeline can be zoomed in and out, to show different levels of detail. As was previously described, Zoetrope can import a number of different features from SOTA multimodal information extraction frameworks, which are displayed in the tracks pane. In its current implementation, Zoetrope has separate tracks for OCRed scene text (Chen et al., 2020) and also for automatically transcribed and POS-tagged spoken language [4]. As an additional textual layer, the tool provides captions from multimodal BLIP embeddings (Li et al., 2022) for all frames. All textual tracks can be searched for keywords in the properties pane. Keyword searches can be verbatim, can use regular expressions or can be more vague, by using FastText embeddings (Bojanowski et al., 2017). For each search, a new track will be automatically generated, indicating where the searched for keyword appears in the video. Besides these basic text search features, Zoetrope also provides information about basic audio features [5], such as an amplitude track that indicates the loudness of the video and also a spectrographic analysis track, indicating which different frequencies are present in the video. A color track shows the mean color of the frames [6] and can be used to detect larger scene structures in the videos. A rather experimental feature of Zoetrope is the dynamics track, which shows the relative difference between CLIP embeddings (Radford et al., 2021) of adjacent frames. As CLIP embeddings are a way to describe what is happening in the video, this is a measure of how much the video content changes in terms of what was shown before.

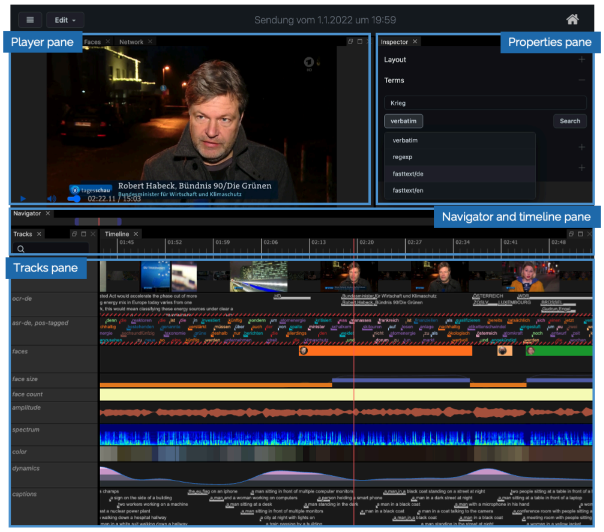


Figure 1: Zoetrope – Interactive feature exploration for a previously selected video.

Finally, Zoetrope provides information on faces (Deng et al., 2019) and face identities (Deng et al., 2021) that were detected in the video. The face size track indicates the size of the largest face on the screen. One possible use of this information is to identify close-ups of single faces in a video. The face count track gives the number of faces visible in the video at a certain time. This can be used to differentiate frames showing 1, 2 or multiple people. The face track clusters similar faces and assigns them to a unique color bar, to indicate where they are actually visible. As for the face information, we are currently also experimenting with additional views (see Fig. 3), such as an overview of all the different clustered faces and their respective emotion values.

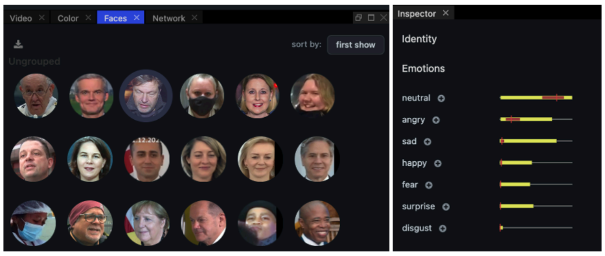


Figure 2: Zoetrope – Clustered faces and face emotion information.

By clicking on a face and selecting different emotion thresholds, tailored analysis tracks can be generated, for instance for all frames that show German politician Robert Habeck with an angry face. As Zoetrope is all about providing different exploratory access points to videos, another experimental visualization is a network view of all faces that co-appear in one frame (see Fig. 4).

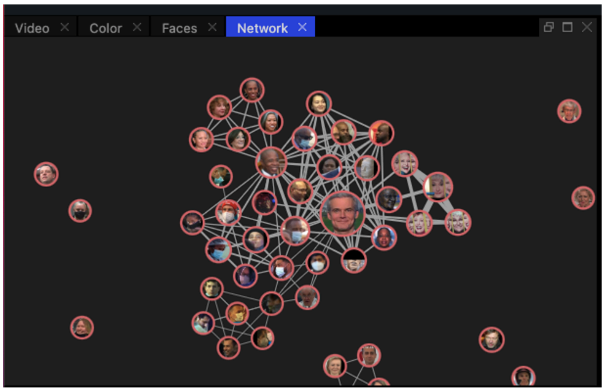


Figure 3: Zoetrope – Network of faces that co-appear in one frame.

Future Directions

Zoetrope is currently being developed as a prototype within the FakeNarratives project, where it is used to support human annotators to identify narrative strategies in news videos. In the future, some of the functions that have proven useful will be transferred to a more sustainable video tool infrastructure that is currently being developed by project partners in Hannover [7]. We hope the prototype will also be interesting for the DH community, as it illustrates how state of the art neural information extraction frameworks can be integrated in an interactive exploratory interface, which could easily be expanded to other areas of multimodal analytics, beyond the current use case of news video analyses. Ultimately, we plan to extend the prototype in the direction of a scalable viewing tool (Burghardt et al., 2018) that can be used to investigate and compare large collections of videos on different levels of detail.

Notes

1. For a comprehensive overview, see Pustu-Iren et al. (2020).
2. FakeNarratives (https://fakenarratives.github.io/) is a collaborative project of the Universities of Bremen, Hannover and Leipzig, generously funded by the German Federal Ministry of Education and Research.
3. Detailed Zoetrope demo video (45:23 minutes), available via https://www.dropbox.com/s/idl0nkenlar3xk7/presentation-zoetrope-2022-11-03.mp4?dl=0
4. We used a pipeline consisting of Mozilla Deepspeech (https://github.com/mozilla/DeepSpeech) and spaCy (https://spacy.io/).
5. We used librosa for audio analysis (https://librosa.org/doc/latest/index.html).
6. We used NumPy for color analysis (https://numpy.org/).
7. “TIB AV Analytics”, more information at https://gepris.dfg.de/gepris/projekt/442397862

Bibliography

Arnold, T., & Tilton, L. (2019). Distant viewing: analyzing large visual corpora. Digital Scholarship in the Humanities, 34(Supplement_1), i3-i16.

Bermeitinger, B., Gassner, S., Handschuh, S., Howanitz, G., Radisch, E., & Rehbein, M. (2019). Deep Watching: Towards New Methods of Analyzing Visual Media in Cultural Studies. Book of Abstracts, DH Conference 2019, Utrecht.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information (arXiv:1607.04606). arXiv. http://arxiv.org/abs/1607.04606

Burghardt, M., Hafner, K. Edel, L., Kenaan, S. & Wolff, C. (2017). An Information System for the Analysis of Color Distributions in MovieBarcodes. In Proceedings of the 15th International Symposium of Information Science (ISI 2017).

Burghardt, M., Kao, M. & Walkowski, NO (2018). Scalable MovieBarcodes – An Exploratory Interface for the Analysis of Movies. 3rd IEEE VIS Workshop on Visualization for the Digital Humanities, Berlin.

Burghardt, M., & Wolff, C. (2014). Humanist-Computer Interaction: Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik. Workshop Proceedings of the DHd working group "Informatik und die Digital Humanities".

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Hrsg.), Computer Vision – ECCV 2020 (Bd. 12346, S. 213–229). Springer International Publishing.

Chen, X., Jin, L., Zhu, Y., Luo, C., & Wang, T. (2020). Text Recognition in the Wild: A Survey (arXiv:2005.03492). arXiv. http://arxiv.org/abs/2005.03492

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. (2019). RetinaFace: Single-stage Dense Face Localisation in the Wild (arXiv:1905.00641). arXiv. http://arxiv.org/abs/1905.00641

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., & Zafeiriou, S. (2021). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2021.3087709

Li, J., Li, D., Xiong, C. & Hoi, S. (2022), „BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". arXiv, 15. Februar 2022. Zugegriffen: 24. Oktober 2022. [Online]. Verfügbar unter: http://arxiv.org/abs/2201.12086

Machill, M., Köhler, S., & Waldhauser, M. (2007). The use of narrative structures in television news: An experiment in innovative forms of journalistic presentation. European Journal of Communication, 22(2), 185-205.

Pustu-Iren, K., Sittel, J., Mauer, R., Bulgakowa, O., & Ewerth, R. (2020). Automated Visual Content Analysis for Film Studies: Current Status and Challenges. DHQ: Digital Humanities Quarterly, 14(4).

Wevers, M., & Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. Digital Scholarship in the Humanities, 35(1), 194-207.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision (arXiv:2103.00020). arXiv. http://arxiv.org/abs/2103.00020

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. Tech. Rep., Technical report, OpenAI.

Tseng, C., Liebl, B., Burghardt, M., & Bateman, J. (2023). FakeNarratives – First Forays in Understanding Narratives of Disinformation in Public and Alternative News Videos. Book of Abstracts, DHd Conference 2023, Trier/Luxemburg.