TIB AV-Analytics: A Web-based Platform for Scholarly Video Analysis and Film Studies

Matthias Springstein
Markos Stamatakis
matthias.springstein@tib.eu
markos.stamatakis@tib.eu
TIB – Leibniz Information Centre for Science and Technology;
L3S Research Center, Leibniz University Hannover
Hannover, Germany

Margret Plank
margret.plank@tib.eu
TIB – Leibniz Information Centre for Science and Technology
Hannover, Germany

Julian Sittel
Roman Mauer
Oksana Bulgakowa
jusittel@uni-mainz.de
romauer@uni-mainz.de
bulgakowa@uni-mainz.de
Johannes Gutenberg University Mainz
Institute of Film, Theatre, Media, and Cultural Studies
Mainz, Germany

Ralph Ewerth
Eric Müller-Budack
ralph.ewerth@tib.eu
eric.mueller@tib.eu
TIB – Leibniz Information Centre for Science and Technology;
L3S Research Center, Leibniz University Hannover
Hannover, Germany

ABSTRACT

Video analysis platforms that integrate automatic solutions for multimedia and information retrieval enable various applications in many disciplines including film and media studies, communication science, and education. However, current platforms for video analysis either focus on manual annotations or include only a few tools for automatic content analysis. In this paper, we present a novel web-based video analysis platform called *TIB AV-Analytics* (*TIB-AV-A*). Unlike previous platforms, *TIB-AV-A* integrates state-of-the-art approaches in the fields of computer vision, audio analysis, and natural language processing for many relevant video analysis tasks. To facilitate future extensions and to ensure interoperability with existing tools, the video analysis approaches are implemented in a plugin structure with appropriate interfaces and import-export functions. *TIB-AV-A* leverages modern web technologies to provide users with a responsive and interactive web interface that enables manual annotation and provides access to powerful deep learning tools without a requirement for specific hardware dependencies. Source code and demo are publicly available at: https://service.tib.eu/tibava.

CCS CONCEPTS

• **Information systems** → *Multimedia and multimodal retrieval*; **Web applications**; • **Human-centered computing** → Information visualization.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591820

KEYWORDS

multimodal video analysis platform, multimedia indexing, computer vision, natural language processing, audio analysis

ACM Reference Format:
Matthias Springstein, Markos Stamatakis, Margret Plank, Julian Sittel, Roman Mauer, Oksana Bulgakowa, Ralph Ewerth, and Eric Müller-Budack. 2023. TIB AV-Analytics: A Web-based Platform for Scholarly Video Analysis and Film Studies. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591820

1 INTRODUCTION

Videos are omnipresent in our everyday lives and serve both to entertain and to provide news and information. They typically consist of different modalities, e.g., image, audio, and text. Methods for automated evaluation of different modalities are crucial to provide researchers with information into the internal dynamics and proportionality of individual videos or even entire corpora. Thus, video analysis platforms that provide researchers from many disciplines, e.g., film and media studies, communication science, sports science, and education, access to state-of-the-art approaches from pattern recognition and multimedia retrieval are of utmost importance to facilitate practical applications in the field of media research.

Several video analysis platforms have been developed to support users in these disciplines. However, while some platforms are limited to manual annotations (e.g., *ANVIL* [10], *Cinemetrics* [23], *ELAN* [24]), other software solutions (e.g., *Videana* [5], *VIAN* [7], *VIAN-DH* [22]) integrate computational approaches to support users in several time-consuming tasks including, for example, shot boundary detection, color analysis, visual concept detection, etc. However, these tools are limited to a few analysis methods and do

not leverage state-of-the-art approaches that have achieved substantial progress in various multimedia tasks in recent years.

In this paper, we present *TIB AV-Analytics* (*TIB-AV-A*) which is a open-source, web-based platform for systematic film and video analysis. It aims to support researchers from various disciplines including film and media studies, communication science, sports science, psychology, and education with state-of-the-art video analysis technologies to which they typically do not have access. For this purpose, we have identified the needs of researchers from these disciplines by conducting a requirement analysis. We also collected feedback from developers and users of other video platforms such as *VIAN* [7] and *ELAN* [24]. To complement existing video analysis platforms, *TIB-AV-A* makes the following contributions. (1) To date, *TIB-AV-A* provides to most complete set of automatic video analysis methods and, unlike existing platforms, integrates state-of-the-art algorithms from computer vision, natural language processing, and audio analysis for automatic extraction of multimodal information from videos. This includes core aspects of video analysis, such as shot boundary detection, automatic speech recognition, shot size classification, as well as further algorithms for face identification, visual concept recognition etc. Moreover, it enables the automatic analysis of audio information including the spoken language. (2) The individual approaches are realized in a plugin structure, which allows developers and researchers to easily integrate plugins to maintain *TIB-AV-A* at the current state of the art. (3) For a web-based, responsive, and interactive experience for users without specific hardware dependencies, *TIB-AV-A* leverages modern web technologies. It supports users to manually add and adjust their own as well as automatically generated annotations. (4) For interoperability with other software and video annotation tools, *TIB-AV-A* allows import and export in common data formats and provides an *Application Programming Interface (API)*.

The remainder of this paper is organized as follows. Section 2 presents the system architecture of *TIB-AV-A*. The video analysis approaches based on recent methods are introduced in Section 3. The current functionalities of *TIB-AV-A* are described in Section 4. Section 5 summarizes this paper and outlines future work.

2 SYSTEM ARCHITECTURE OF TIB-AV-A

TIB-AV-A offers a wide range of state-of-the-art video analysis methods and manual annotation of video segments. To simplify extensions, the individual components were built as modular as possible with plugins. As shown in Figure 1, *TIB-AV-A* consists of four main components: frontend, backend, analyzer, and inference server.

2.1 Frontend

The frontend allows individual users to upload videos, configure and launch processing pipelines, and annotate video segments. The frontend was implemented using the *Vue.js* framework and the *WebGL* library *PixiJS*. *PixiJS* enables *Graphics Processing Unit (GPU)* acceleration and allows a more interactive use of the tool especially when many visualization for a video are displayed.

2.2 Backend

The task of the backend is to store and manage user inputs including uploaded videos and manually created annotations. Furthermore, it defines processing pipelines that the user can execute. A processing pipeline consists of one or more modular analyzer plugins (Section 2.3) to perform a certain video analysis task. For example, the extraction of the audio spectrogram (Section 3.1) performs two plugins to (1) extract the audio signal from the video, and (2) calculate its frequency histogram. The backend was implemented using the *DJango* framework and communicates with the frontend using a *Representational State Transfer (REST) API* and with the analyzer using *gRPC Remote Procedure Calls (gRPC)* protocol. The *REST API* allows external developers to use *TIB-AV-A*'s plugins. Users as well as their annotations and pipeline results are stored in a *Structured Query Language (SQL)* database.

2.3 Analyzer

The analyzer is a *gRPC* server implemented in python and serves as a job queue for plugins. To simultaneously process multiple requests from different users, a job scheduler is added, which processes incoming calls from plugins with multiple workers. The individual workers provide the current status, which gives users feedback on the plugin's progress. Furthermore, the analyzer includes a cache that stores individual plugin results.

2.4 Inference

Current approaches for the analysis of multimedia content are implemented with different frameworks, such as *TensorFlow* [1] or *PyTorch* [14]. We use a *BentoML* inference server with *REST API* to support simultaneous use of various versions of machine learning frameworks. This simplifies the implementation of new plugins in *TIB-AV-A*. In addition, the inference server automatically manages the available computational resources such as accelerators (*GPU*).

3 VIDEO CONTENT ANALYSIS

In this section, we describe the algorithms used in *TIB-AV-A* for the analysis of the auditory (Section 3.1) and visual content (Section 3.2) of the video. The currently implemented algorithms were chosen based on their tradeoff in speed, performance, and availability (license). Section 3.3 presents details on how users can combine outputs of different analysis methods to extract high-level features using logical operators. The selection of approaches was determined and prioritized based on a requirements analysis conducted with users from various disciplines, including communication science and film studies. Surprisingly, potential users were also interested in low-level features from audio (e.g., frequency and volume) and image (e.g., color and brightness) besides more complex information extraction techniques. To date, we mainly focused on the integration of approaches for automatic visual content analysis. However, first approaches to analyse the audio and speech content have been already implemented and will be extended in the future.

3.1 Audio & Speech Content Analysis

Low-level audio features, such as the dynamics, volume, and spectrum of the audio signal, can provide information about various aspects of audio analysis, such as the identification of music, speech, etc. We use the *librosa* [12] python library to extract low-level audio features from videos including the amplitude curve (waveform), volume (root mean square), and the frequency spectrogram.

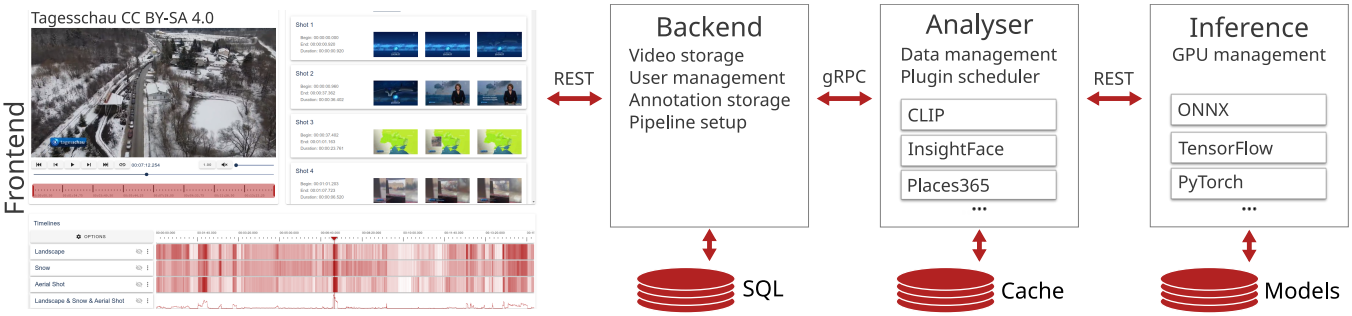

Figure 1: System architecture of *TIB-AV-A* with associated interfaces and database structure.

The speech in videos contains many important information including topics, entities, and sentiment. We apply *Whisper* [17], a state-of-the-art transformer for automatic speech recognition, to automatically extract the speech transcripts from the videos. The transcripts are an important foundation for many subsequent speech analysis steps using natural language processing approaches, e.g, for named entity linking [11], sentiment analysis [18], and topic modeling [2], that will be implemented in *TIB-AV-A* in the future.

3.2 Visual Content Analysis

3.2.1 Low-level Visual Features. Low-level visual features such as color and brightness can provide users with valuable insights on the video content. We have used the python library *scikit-learn* [15] to cluster the RGB values of all video pixels within a frame using the k-means clustering approach. The number of clusters k can be defined by the user in the frontend. In addition, we averaged the value (V) in the HSV color space to extract the brightness.

3.2.2 Basic Video Analysis. The detection of shot boundaries is one of the most fundamental task in video analysis. We apply *TransNet V2* [21] for shot boundary detection as it provides robust and accurate results. Furthermore, the shot boundaries were used for a *Kernel Density Estimation* using *scikit-learn* [15] to compute the *cut density* [19], i.e., the frequency of shot transitions, that helps spot video segments with a low or high shot frequency (e.g., actions scenes). Another important aspect according to our requirement analysis is the shot size (e.g., close-up, long shot). Since there are no publicly available models, we trained a *ResNet-50* [8] on the *MovieNet* dataset [9] that is annotated for five shot sizes.

3.2.3 Visual Concept Detection & Face Analysis. The detection of visual concepts, such as places, objects, persons, and actions was identified as another important aspect for film and media studies. To predict the environmental setting of a shot, we apply a pretrained ResNet-50 [8] model trained on the *Places365* dataset [25]. To detect and identify faces, we use the *insightface* python library that offers a variety of models for face detection [6] and facial feature representation [3]. The identification of persons is realized by a comparison of the facial features between the faces found within a video to the face depicting the queried person in an example image which is provided from the user via the frontend. To extract the facial emotions of each face found in the video, we apply the *facial attribute analysis* from *deepface* [20].

3.2.4 Zero-shot Concept Classification. The set of relevant concepts depicted in videos highly depends on various aspects, such as the application domain (e.g., news analysis, film studies, etc.), topic or genre, and the user's interest. Thus, we apply recent approaches on visual representation learning based on vision-language pretraining [13, 16] that, unlike previous methods that rely on a fixed set of discrete labels [4, 8], allow us to automatically classify *arbitrary task-relevant* visual concepts in videos based on textual descriptions (also referred to as *prompting* [26]). To describe individual frames, we use *CLIP (Contrastive Language-Image Pretraining)* [16] as it has proven to be very effective for many downstream tasks. Furthermore, we apply *X-CLIP* [13] to enable a zero-shot visual concept detection for tasks that benefit from the temporal context in the video, such as the recognition of (inter)actions and gestures.

3.3 Logical Operators for Pattern Recognition

We collaborated closely with researchers in the communication science and film studies to develop *TIB-AV-A*. We noticed that researchers are often interested in patterns that comprise a combination of features. For example, they are interested in analysing specific movie or news pieces that contain a *close-up* of a *person* with a *sad facial expression* or an *aerial shot* of a *landscape* covered in *snow* (Figure 1). It can be also used to add conditions to certain patterns to, for example, search for specific actions if a person visible. For this purpose, we implemented a plugin that combines the probabilities of certain features (e.g., scenes, emotions, shot sizes) with logical operations. So far, user can define high-level features by aggregating them with the logical *or* and the logical *and* function.

4 DEMONSTRATOR

TIB-AV-A is available on: https://service.tib.eu/tibava. Users can add videos to their collection by uploading a video file and its meta information. Once a video has been uploaded, the video analysis view depicted in Figure 2 is presented. It is composed of four main components: (a) The video and its control elements (e.g., audio volume, playback speed, timeline navigation) are displayed in the top left corner. (b) In the top right corner, the user is provided with a list of shots (b1) and annotations (b2) of the current frame. (c) In the bottom of the screen, the timelines containing manual annotations and plugin results including automatically generated annotations as well as numerical results are visualized. (d) In the navigation

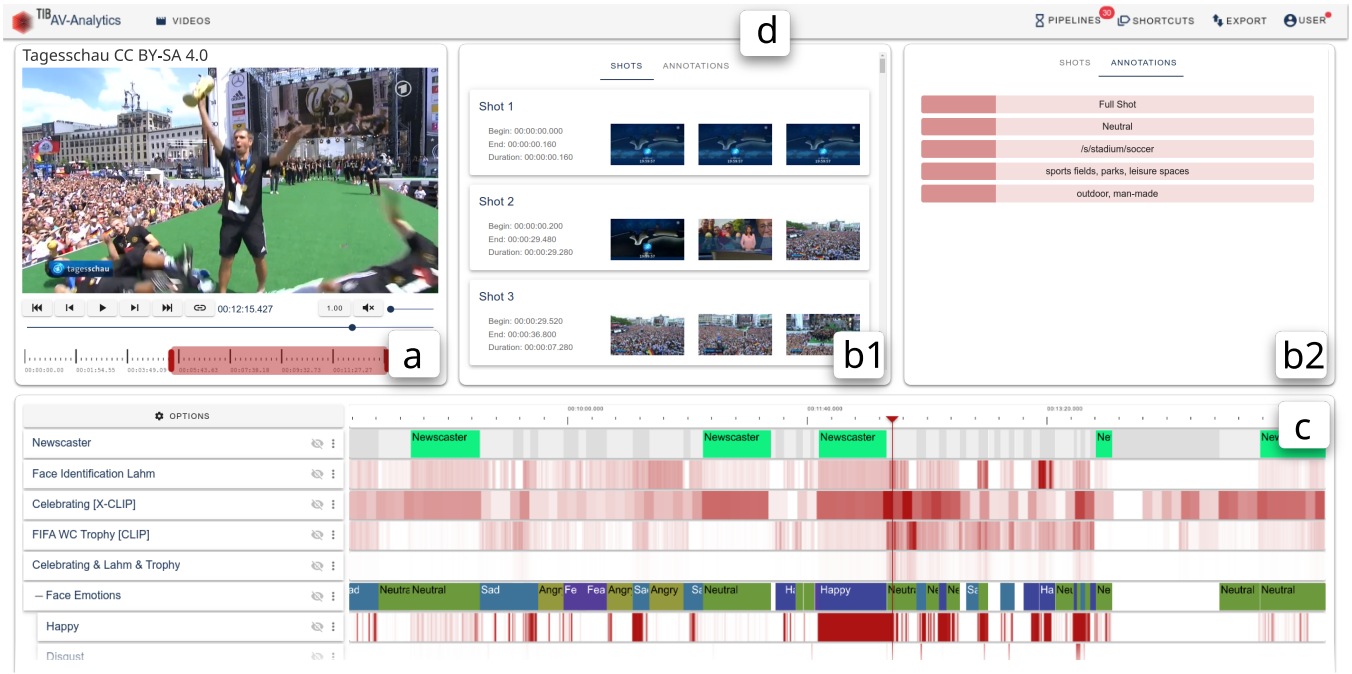

Figure 2: Interface of *TIB-AV-A* for a German news broadcast. It contains a video player (a), overview of detected shots (b1) and annotations (b2), timelines (c) that can display categorical (e.g., *Newscaster*) and numerical values (e.g., *FIFA WC Trophy [CLIP]*), and a navigation bar (d). Timelines with numerical values indicate, e.g., the probability whether a concept is depicted in a video. The user can select the visualization type (line chart, color chart) and color (here: from white unlikely to red likely).

bar users can start pipelines, add shortcuts for manual annotations, export results, and get access to the user management.

Users have access to all video content analysis steps explained in Section 3. For each uploaded video, shot boundary detection is automatically performed as it serves as a basis for many subsequent video analysis steps and manual annotations. The shot boundary detection creates a timeline with timeline segments that correspond to the individual shots. The parameters, such as the the number of frames per second for which the pipeline is performed, can be configured in the pipeline menu. The user can monitor the progress of running pipelines in the navigation bar. Once a pipeline has finished, the corresponding timeline will be displayed. To allow customization, different visualization types (e.g., line chart, heatmaps) and color schemes can be selected. Timelines can be (or are already) grouped and arranged hierarchically, e.g, the *shot sizes* contain five different classes. These groups can be expanded and collapsed. Moreover, timelines can be rearranged, deleted, and duplicated.

Besides automatic video analysis, *TIB-AV-A* also allows users to add and change annotations by clicking on an annotation segment, i.e., shot, or selecting a time range. In the shortcut menu, users can assign keyboard shortcuts for existing annotation labels. Moreover, users can navigate timeline segments using the arrow keys on the keyboard for efficient annotation and can split or merge them to modify the shot boundaries. In the option menu within the timelines view, users are also presented with the option to import and export timelines. Currently, we support import and export to *ELAN* [24] as well as the export as csv (comma-separated values) file.

5 CONCLUSIONS

In this paper, we have presented a novel video analysis platform called *TIB-AV-A*. Unlike exiting platforms, *TIB-AV-A* uses modern web technologies and integrates state-of-the-art methods for automatic video analysis. To date, it offers the most complete set of analysis tools within a video analysis platform to support researchers in many disciplines, e.g., film and media studies, communication science, and education. Moreover, *TIB-AV-A* allows manual annotations and provides interoperability to existing video analysis efforts through import and export in common data formats and an API. The plugins are developed in a modular fashion to simplify future extension of *TIB-AV-A*. In the future, we want to provide users with the possibility to create custom data visualizations directly within *TIB-AV-A*, e.g., line charts, bar charts, and graphs that indicate the (co-)occurrence of persons or the places specific persons or concepts appear in. Furthermore, we aim to add plugins for other important use cases, e.g., natural language processing approaches to analyze the written and spoken text. Although *TIB-AV-A* was developed to analyze individual videos, we plan to extend the platform to allow information retrieval in a video corpus.

ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG) under the project number 442397862. We would like to thank Jim Rhotert and Junaid Ghauri (both TIB – Leibniz Information Centre for Science and Technology) for their help in developing *TIB-AV-A*.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint* abs/1603.04467 (2016). arXiv:1603.04467 http://arxiv.org/abs/1603.04467

[2] Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, Virtual Event, August 1-6, 2021*. Association for Computational Linguistics, 759–766. https://doi.org/10.18653/v1/2021.acl-short.96

[3] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. 2020. Sub-center ArcFace: Boosting Face Recognition by Large-Scale Noisy Web Faces. In *European Conference on Computer Vision, ECCV 2020, Glasgow, UK, August 23-28, 2020*. Springer, 741–757. https://doi.org/10.1007/978-3-030-58621-8_43

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[5] Ralph Ewerth, Markus Mühling, Thilo Stadelmann, Julinda Gllavata, Manfred Grauer, and Bernd Freisleben. 2009. Videana: A Software Toolkit for Scientific Film Studies. *Digital Tools in Media Studies – Analysis and Research, An Overview* (2009), 101–116.

[6] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. 2022. Sample and Computation Redistribution for Efficient Face Detection. In *International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=RhB1AdoFfGE

[7] Gaudenz Halter, Rafael Ballester-Ripoll, Barbara Flückiger, and Renato Pajarola. 2019. VIAN: A Visual Annotation Tool for Film Analysis. *Computer Graphics Forum* 38, 3 (2019), 119–129. https://doi.org/10.1111/cgf.13676

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[9] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. MovieNet: A Holistic Dataset for Movie Understanding. In *European Conference on Computer Vision, ECCV 2020, Glasgow, UK, August 23-28, 2020*. Springer, 709–727. https://doi.org/10.1007/978-3-030-58548-8_41

[10] Michael Kipp. 2014. ANVIL: A Universal Video Research Tool. *Handbook of Corpus Phonology* (2014), 420–436.

[11] Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient One-Pass End-to-End Entity Linking for Questions. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 6433–6441. https://doi.org/10.18653/v1/2020.emnlp-main.522

[12] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Python in Science Conference*, Vol. 8.

[13] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding Language-Image Pretrained Models for General Video Recognition. In *European Conference on Computer Vision, ECCV 2022, Tel Aviv, Israel, October 23-27, 2022*. Springer, 1–18. https://doi.org/10.1007/978-3-031-19772-7_1

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. 8024–8035. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[15] 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.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[17] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint* abs/2212.04356 (2022). https://doi.org/10.48550/arXiv.2212.04356 arXiv:2212.04356 Whisper.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[19] Nick Redfern. 2022. Analysing motion picture cutting rates. *Wide Screen* 9, 1 (2022), 1–29.

[20] Sefik Ilkin Serengil and Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *International Conference on Engineering and Emerging Technologies, ICEET 2021, Istanbul, Turkey, October 27-28, 2021*. IEEE, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697

[21] Tomás Soucek and Jakub Lokoc. 2020. TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv preprint* abs/2008.04838 (2020). arXiv:2008.04838 https://arxiv.org/abs/2008.04838

[22] Pascal Forny Teodora Vuković, Christoph Hottiger. [n. d.]. *VIAN-DH*. Retrieved Feb 21, 2022 from https://www.liri.uzh.ch/en/projects/VIAN-DH.html

[23] Yuri Tsivian. [n. d.]. *Cinemetrics*. Retrieved Feb 21, 2022 from http://www.cinemetrics.lv

[24] Peter Wittenburg, Hennie Brugman, Albert Russel, Alexander Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006*. European Language Resources Association (ELRA), 1556–1559. http://www.lrec-conf.org/proceedings/lrec2006/summaries/153.html

[25] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

[26] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348. https://doi.org/10.1007/s11263-022-01653-1