Encoding and Decoding Narratives: Datafication and Alternative Access Models for Audiovisual Archives

Yuchen Yang
Laboratory for Experimental Museology, EPFL
Lausanne, Switzerland

ABSTRACT

Situated in the intersection of audiovisual archives, computational methods, and immersive interactions, this work probes the increasingly important accessibility issues from a two-fold approach. Firstly, the work proposes an ontological data model to handle complex descriptors (metadata, feature vectors, etc.) with regard to user interactions. Secondly, this work examines text-to-video retrieval from an implementation perspective by proposing a classifier-enhanced workflow to deal with complex and hybrid queries and a training data augmentation workflow to improve performance. This work serves as the foundation for experimenting with novel public-facing access models to large audiovisual archives.

CCS CONCEPTS

• **Applied computing** → **Arts and humanities**; • **Information systems** → **Multimedia information systems**; **Information retrieval**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

text-to-video encoding, computational archive, experimental museology, audiovisual archive

ACM Reference Format:
Yuchen Yang. 2023. Encoding and Decoding Narratives: Datafication and Alternative Access Models for Audiovisual Archives. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3581783.3613434

1 INTRODUCTION

This work focuses on real-world archives - Télévision Suisse Romande (RTS) - and aims to tackle the preservation and accessibility issue in the age of data. The work is divided into two intertwined sub-parts:

A model for datafication. This part emphasises the preservation end for archives from a data perspective. The first goal is to propose an ontology to formally represent various levels and aspects of data from AV archives and interactive experiences. The proposed ontological model reflects the data lineage among technical, content, conceptual, and interaction (meta)data and nurtures an update of preservation practices from a data management perspective. Using the model as a starting point, this part of the research will propose a data system and schema addressing retrieval efficiency and effectiveness. Together, deliverables from this section will provide theoretical and practical support for the preservation and new explorative methods.

Encoding AV archive: text-to-video embedding. This part of the research focuses on solving accessibility issues by verifying and improving the state-of-the-art text-to-video methods for real-world archives. The objective for this part is to explore the usage of such novel models beyond standard datasets and search for specific videos, Fusing it into various workflows creates applicable solutions to support real-world applications that focus on meaningful explorations.

2 RELATED WORK

2.1 Archives and datafication

In recent years, AV archives started to experiment with a deeper operationalisation on the content level, building manual or automated tools for annotations [3], and finding new ways of using content descriptors and the ever-growing surrounding knowledge. Some focus on themed analysis - using colour to analyse aesthetics [16] or movements for choreography [6]. More general ones focus on tool sets for semantic annotations for audiovisual content [10, 41].

With more diverse data, curatorial practices with AV archives shift to immersive and interactive experiences to explore the plurality of memory materials and encourage personalised sense-making [23]. The Pods in the Eye Filmmuseum [20] aims to explore the remixing of historical archives and the SEMIA project [27] works on experimenting with alternative archive interfaces. Commercial tools like the Storyformer [38] also become available for creating personalised and interactive content experiences.

However, such fragmented practices remain impulsive and unsystematic. Few seek to reconcile the content, technology, and curatorial needs into reusable solutions and there is a lacking of fundamental models for mapping, linking, and managing the growing complex data.

2.2 text-to-video embedding

The recently popularised text-to-video retrieval task takes an arbitrary text query and searches for the most relevant video clips accordingly. First proposed in 2016 [33], the text-to-video embedding usually has a two-stream architecture [28, 29] utilising videos and corresponding descriptions. Videos and texts are transformed by encoders into vectors and projected into a common feature space. The paired relationships of the vectors are then used for training

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3613434

(or transforming) this common feature space so that the paired vectors are the closest to each other. This trained common feature space, called joint embedding space, is then used for inference and retrieval. Bi-directional loss [22], symmetric cross entropy loss[40], and triplet loss [36] are often used to train a joint embedding space. The training of text-to-video retrieval models benefits from manually labelled datasets for language-to-video related tasks [1, 33, 42].

Some works focus on improving the encoding of video using multimodal cues such as the face, audio, and speech [19, 37], while others work on generalising such models with large-scale pretraining. [28] uses machine-generated transcription as descriptions for videos to construct large training sets. [14, 25, 30, 43] rely on large language to image models to pretrain, and finetune the model with text-video datasets.

While methods are maturing, the ability to work with real-world archives and queries is unknown. Models' performance drops on datasets with sophisticated video descriptions [21]. A recent study has proven that improved annotations would improve models' performance [8].

3 METHODS AND EXPERIMENTS

3.1 A model for datafication

Based on existing works [12, 26, 34], a mapping of the complexity of AV archive metadata in various dimensions is made and summarised in Figure 1.

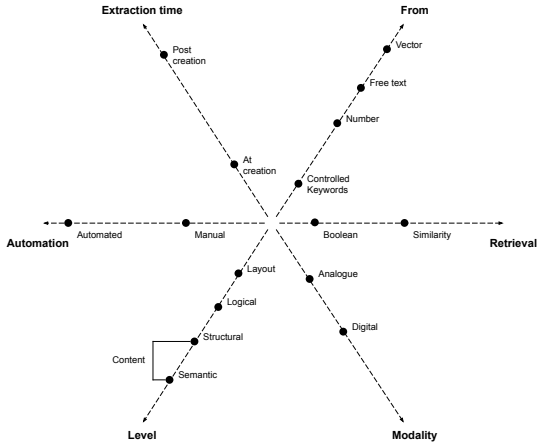

Figure 1: Dimensions of the metadata of a multimedia object visualised in the level, the automation, the extraction time, the form, the retrieval, and the modality dimensions

Schemas like MPEG-7 [26] are often used for such complexity in multimedia data and metadata management. However, they are not entirely suitable for content-level descriptors nowadays. Such schemas believe low-level metadata can be extracted automatically (like colour, texture, shape, timbre, pitch and rhythm), and high-level metadata are human annotations on a conceptual level (like emotions, content summaries) [35]. However, With maturing machine learning methods, the semantic gap of the inability to create high-level conceptual descriptors in an automated way is closing down [39]. Current standards have not reflected such changes. For example, descriptors such as body key points, text-to-video embedding vectors, and metadata (of training data, model, etc.) for machine learning methods are not considered at all. These are essential for understanding, improving, and managing the results, as well as a successful modern data management system [18, 24].

On the other hand, with archival practices increasingly adapting interactive and personalised elements, only considering the content end is not enough. The public value of archives is realised through the combination of the **Content**, **Participants**, and **Interactions** [4, 7, 13]. However, the connection and relationship between these three parts have never been formalised. Inspired by previous works to formalise textual narratives using ontology [5] and formalise the composition and relationship of multimedia big data [32], this part of the research aims to provide an ontological model to describe the interconnection between the three aspects.

3.2 Encoding AV archive: text-to-video embedding

3.2.1 Base text-to-video embedding model. Since the purpose is to evaluate and improve the performance of text-to-video models when dealing with queries beyond plain visual description (such as with speech info). Models that do not consider audio information are discarded. Models are selected following these principles: representativeness, source code availability, and performance ranking. In the end, this work chooses the classic multimodal model MTT [19] and the latest multimodal retrieval model MFT [37] as the core to construct the two proposed workflows.

3.2.2 Base dataset. This work uses the standard dataset MSR-VTT [42] as a base, which contains videos in music, sports, news, movie, drama, etc. The vastly diversified content mirrors the RTS archives the most. This dataset provides 10,000 video clips harvested from random internet sites, totalling 41.2 hours. Each video clip within this dataset is paired with 20 human annotations, contributing to 200,000 clip-sentence pairs.

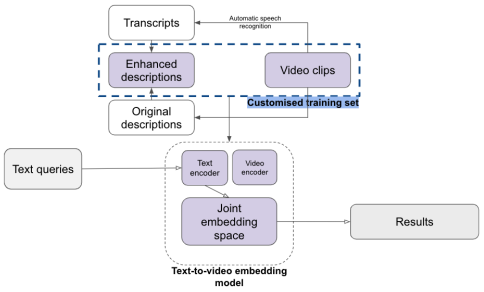

Figure 2: A detailed look at the proposed single model with enhanced description workflow, the key component - customised training set - of this workflow is highlighted in the blue dashed box.

3.2.3 Single model with enhanced description workflow. **Customised training set.** Fig.2 depicts the workflow from end to end. The original annotations in the MSR-VTT are limited to plain and visual descriptions. For a video of a conversation between a contestant and judges, the annotations are "a girl and the judges talking on the voice" and "a girl is talking to the judges on a game show", which do not reflect the content of the conversation at all. To better understand if improving annotations would help with model performance, this workflow focuses on building a customised training set with the enhanced descriptions following a previous work using automatic speech recognition (ASR) [29]. The customised training set is contracted by randomly replacing one or more of the 20 original descriptions paired with a video clip with ASR results.

Workflow implementation details. Transcripts. The MSR-VTT dataset is fed to Whisper [31] to obtain transcripts, using the provided "small model" and following the huggingface guides[1]. **Customised training set.** We use 1k-A split on MSR-VTT produced by [44] for constructing the customised training set. **Model Training.** The customised training set is then used to train the two base models, MMT and MFT, following their official implementation configurations respectively[2]. This workflow utilises a customised training set and produces two new methods: customised MMT and customised MFT.

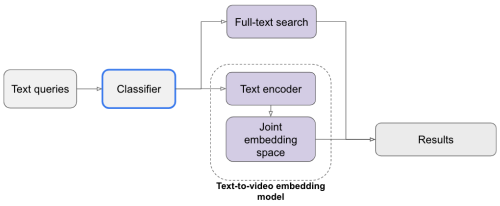

Figure 3: A detailed look at the proposed classifier-enhanced workflow, the key component - classifier - of this workflow is highlighted in the blue box.

3.2.4 Classifier-enhanced workflow. **Classifier.** As seen in Fig.3, the fundamental idea is to classify queries and send them to the appropriate retrieval component. Although hard-coded rule sets can function to a certain extent, a machine-learning-based classifier is introduced in the hope of scalability and generalisation. In this work, the classifier distinguishes quote or speech-related texts from plain visual descriptions. Recent advancements in sequence models have brought significant transformations in natural language processing (NLP). Recurrent neural networks (RNNs) and transformers have consistently performed remarkably on numerous standard NLP benchmarks. We employ a long short-term memory (LSTM) architecture in this workflow to address the classification task.

Workflow implementation details. Training data for the classifier. A labelled dataset with speeches or quotes and plain visual descriptions is prepared for the training. The labelled data for speeches or quotes are constructed by joining 2000 sentences from online databases[3] using the regular expression filters converted advisors [4], and 1000 transcripts from the customised training set in Section 3.3.1. The labelled plain visual descriptions are 2000 random descriptions from the MSR-VTT 1k-A training set. An 80-20 split is done for training and testing. **Training the classifier.** The binary classifier is trained following previous work[5] with the TensorFlow Keras sequential model. It contains an embedding layer representing each word with a vector length of 16. The following 16-unit LSTM layer uses relu activation. The final dense layer has seven units and a softmax activation for classification. The model is fit on the training dataset with a batch size of 32 for seven epochs. **Text-to-video model.** This workflow sends non-speech or non-quote queries to the text-to-video model. The original MSR-VTT 1k-A training set is used to train the two base text-to-video embedding models MMT and MFT, following their official implementation configurations respectively[6]. **Full-text search model.** For quote- or speech-related queries, the workflow sends them to the full-text search. The ASR results from MSR-VTT in Section 3.3.1 are stored in an ElasticSearch[7] database. The similarity is calculated by built-in API for full-text queries. This workflow utilises a classifier and produces two new methods based on the two base text-to-video embedding models: classifier MMT and classifier MFT.

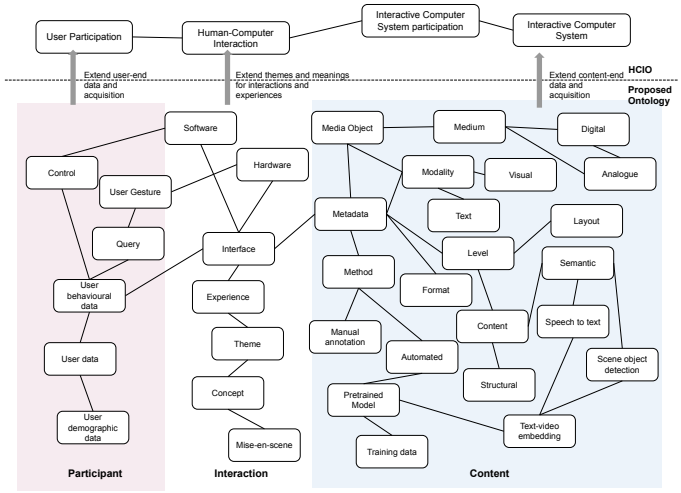

Figure 4: An exemplary look at some of its key classes and properties of it to illustrate the overall ontology concept

4 PRELIMINARY RESULTS

4.1 A model for datafication

This ontology is modulated along three corresponding dimensions - content, participant, and interaction (Fig. 4) The proposed ontology

[1]https://huggingface.co/openai/whisper-small
[2]MFT:https://github.com/ninatu/everything_at_once; MMT: https://github.com/gabeur/mmt
[3]https://libguides.bgsu.edu/c.php?g=227160&p=1505718
[4]https://www.ccis.edu/student-life/advising-tutoring/writing-math-tutoring/introduce-quotations
[5]https://medium.com/holler-developers/intent-detection-using-sequence-models-ddae9cd861ee
[6]MFT:https://github.com/ninatu/everything_at_once; MMT: https://github.com/gabeur/mmt
[7]https://www.elastic.co/

uses HCIO ([11]) as a core and top-level reference for inheriting fundamental notions for describing the core aspects of the human-computer interaction phenomenon. Ontology for Media Resources[8], COMM ([2]), and an ontology for harvesting user input in the immersive environment ([9]) are used as starting point to populate the content, participant, and interaction dimensions.

4.2 Encoding AV archive: text-to-video embedding

4.2.1 Test setup. **Baseline test set.** The popular 1k-A split on MSR-VTT provides the baseline test dataset. **Customised test set.** A customised test set is introduced to evaluate different methods' performance in complex query situations better. The customised set is constructed on the base of the 1k-A split test set on MSR-VTT. We randomly replaced 50% of the original ground truth (one random entry from the 20 annotations) with the obtained transcripts for those video clips. The result is 1,000 ground truth pairs with a mixture of plain visual descriptions and speeches or quotes. **Evaluation metrics.** Standard matrics R@5 are likely to be more useful when understanding the performance and hence picked to report in the result.

4.2.2 Comparison with state-of-the-art. Table 1 reports the result of all methods constructed for the evaluation on the comparison with the state-of-the-art. The baseline models' performance on the original MSR-VTT 1k-A train-test split is referenced when available from the original paper. Following the official implementations, the two models are also trained according to the original training settings from scratch using the 1k-A split of the original and the customised MSR-VTT dataset. The test sets from the 1k-A split of the original and the customised MSR-VTT dataset are used to conduct the final evaluation. Overall, the proposed Classifier MFT method achieved comparable performance with the baseline state-of-the-art model, with an R@5 of 54.2 compared to 57.1. On the customised MSR-VTT test set, where the query situation is a bit more complex, Classifier MFT outperforms all other methods with an R@5 of 77.5.

Method	Training Dataset	Original MSR-VTT R@5↑	Customised MSR-VTT R@5↑
MMT	Original MSR-VTT	54.0	12.9
MFT	Original MSR-VTT	**57.1**	11.2
Customised MMT	Customised MSR-VTT	49.5	19.0
Customised MFT	Customised MSR-VTT	47.0	**22.1**
Classifier MMT	Original MSR-VTT	52.7	76.2
Classifier MFT	Original MSR-VTT	**54.2**	**77.5**

Table 1: Results of the baseline, customised, and classifier-enhanced method on the original and customised MSR-VTT test sets.

Several observations can be made based on the experiment results. First, all four baseline and customised methods suffer a performance drop when dealing with quote-related queries specific to speech information. This expected behaviour could be caused by the fact that most of the speech information is not matched with the visual perspective of the video clips in the given dataset. Second, models trained on the customised dataset, which contains descriptions that are transcriptions of the video, perform slightly worse when tested with the original test set, but better when dealing with a hybrid of descriptive and quote-related queries. Adding the speech-related descriptions in the customised dataset can provide more information during the training and slightly improve the performance when dealing with queries targeted more on the audio perspective. However, the extra information can also be regarded as noise, messing up the joint-embedding space and undermining the overall performance. Third, all classifier-enhanced methods perform well in both test sets. However, it is noticeable that the performance when dealing with the original test set drops slightly compared to the baseline methods. This can be caused by the fact that the performance is heavily determined by the classifier's performance, in which case it will not be 100% accurate.

[8]https://www.w3.org/TR/mediaont-10

5 FUTURE WORK

A model for datafication The development of this ontology is an ongoing and circular task. Taking advantage of the Sinergia project, this ontology will be developed, validated, and improved with archival partners and exhibition installation. Another natural next step is to build a flexible data schema (JSON) to store the array of feature vectors and metadata with temporospatial consideration to facilitate better retrieval or explorative applications.

Encoding AV archive: text-to-video embedding Multiple works on text-to-video retrieval methods have emphasised the importance of having a more situated and better quality dataset in improving the retrieval performance [8, 15, 37]. In this specific work, only one additional quote-related query text is considered. However, narrative text, even as simple as a diary, has a much more diverse type of sentence describing many different aspects and levels of semantics within AV content. It would be beneficial to dig deeper in that direction and find a better strategy to create an appropriate customised annotation for video clips to reflect that. For instance, Vlogs, with the speech information being diverse enough to include many aspects of the given video, could be a more suitable source of descriptions for creating the text-video pairs to cover a more diverse scenario in the query [17]. However, if the more complex annotation will be regarded as noise and hinder the performance is yet to be tested. New training strategies and architectures to handle the weakly-paired text are also required.

Together, the two parts of this research serve as the foundation and necessary support for exploring new access models for large AV archives from a novel and data-driven perspective. On top of this basis, the thesis will move on to build and evaluate prototypes for innovative experiences featuring more personalized, intuitive, serendipitous, and human-centric explorations of AV archives.

ACKNOWLEDGMENTS

This research is made possible through the SNSF Sinergia project, grant number CRSII5_198632.

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*. 5803–5812.

[2] Richard Arndt, Raphaël Troncy, Steffen Staab, and Lynda Hardman. 2009. Comm: A core ontology for multimediaannotation. *Handbook on Ontologies* (2009), 403–421.

[3] Taylor Arnold, Stefania Scagliola, Lauren Tilton, and Jasmijn Van Gorp. 2021. Introduction: Special Issue on AudioVisual Data in DH. *DHQ: Digital Humanities Quarterly* 15, 1 (2021).

[4] Tricia Austin. 2020. *Narrative environments and experience design: Space as a medium of communication.* Routledge.

[5] Valentina Bartalesi, Carlo Meghini, and Daniele Metilli. 2016. Steps towards a formal ontology of narratives based on narratology. In *7th Workshop on Computational Models of Narrative (CMN 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[6] Peter Broadwell and Timothy R Tangherlini. 2021. Comparative K-Pop Choreography Analysis through Deep-Learning Pose Estimation across a Large Video Corpus. *DHQ: Digital Humanities Quarterly* 15, 1 (2021).

[7] Danielle Carter. 2018. Narrative learning as theory and method in arts and museum education. *Studies in Art Education* 59, 2 (2018), 126–144.

[8] Haoran Chen, Jianmin Li, Simone Frintrop, and Xiaolin Hu. 2022. The MSR-Video to Text dataset with clean annotations. *Computer Vision and Image Understanding* 225 (2022), 103581.

[9] Catalin-Marian Chera, Wei-Tek Tsai, and Radu-Daniel Vatavu. 2012. Gesture ontology for informing service-oriented architecture. In *2012 IEEE International Symposium on Intelligent Control*. IEEE, 1184–1189.

[10] Allison Cooper, Fernando Nascimento, and David Francis. 2021. Exploring Film Language with a Digital Analysis Tool: The Case of Kinolab. *DHQ: Digital Humanities Quarterly* 15, 1 (2021), 1–29.

[11] Simone Dornelas Costa, Monalessa Perini Barcellos, Ricardo de Almeida Falbo, Tayana Conte, and Káthia M de Oliveira. 2022. A core ontology on the Human–Computer Interaction phenomenon. *Data & Knowledge Engineering* 138 (2022), 101977.

[12] Arjen P de Vries and HM Blanken. 1998. Database technology and the management of multimedia data in the Mirror project. In *Multimedia Storage and Archiving Systems III*, Vol. 3527. SPIE, 443–453.

[13] EVİNÇ Dogan and M Hamdi KAN. 2020. Bringing heritage sites to life for visitors: towards a conceptual framework for immersive experience. *Advances in Hospitality and Tourism Research (AHTR)* (2020), 1–24.

[14] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3354–3363.

[15] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097* (2021).

[16] Barbara Flueckiger and Gaudenz Halter. 2020. Methods and Advanced Tools for the Analysis of Film Colors in Digital Humanities. *DHQ: Digital Humanities Quarterly* 14, 4 (2020).

[17] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. 2018. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4991–5000.

[18] Manel Fourati, Anis Jedidi, and Faiez Gargouri. 2020. A survey on description and modeling of audiovisual documents. *Multimedia Tools and Applications* 79 (2020), 33519–33546.

[19] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 214–229.

[20] Grazia Ingravalle. 2015. Remixing early cinema: historical explorations at the EYE Film Institute Netherlands. *Moving Image: the Journal of the Association of Moving Image Archivists* 15, 2 (2015), 82–97.

[21] Kaixiang Ji, Jiajia Liu, Weixiang Hong, Liheng Zhong, Jian Wang, Jingdong Chen, and Wei Chu. 2022. Cret: Cross-modal retrieval transformer for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 949–959.

[22] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems* 27 (2014).

[23] Sarah Kenderdine, Ingrid Mason, and Lily Hibberd. 2021. Computational Archives for Experimental Museology. In *International Conference on Emerging Technologies and the Digital Transformation of Museums and Heritage Sites*. Springer, 3–18.

[24] Arun Kumar, Matthias Boehm, and Jun Yang. 2017. Data management in machine learning: Challenges, techniques, and systems. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 1717–1722.

[25] Alexander Kunitsyn, Maksim Kalashnikov, Maksim Dzabraev, and Andrei Ivaniuta. 2022. Mdmmt-2: Multidomain multimodal transformer for video retrieval, one more step towards generalization. *arXiv preprint arXiv:2203.07086* (2022).

[26] Bangalore S Manjunath, Philippe Salembier, and Thomas Sikora. 2002. *Introduction to MPEG-7: multimedia content description interface*. John Wiley & Sons.

[27] Eef Masson, Christian Gosvig Olesen, Nanne van Noord, and Giovanna Fossati. 2020. Exploring Digitised Moving Image Collections: The SEMIA Project, Visual Analysis and the Turn to Abstraction. *DHQ: Digital Humanities Quarterly* 4 (2020).

[28] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9879–9889.

[29] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2630–2640.

[30] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. 2021. A straightforward framework for video retrieval using clip. In *Pattern Recognition: 13th Mexican Conference, MCPR 2021, Mexico City, Mexico, June 23–26, 2021, Proceedings*. Springer, 3–12.

[31] 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 arXiv:2212.04356* (2022).

[32] Antonio M Rinaldi and Cristiano Russo. 2018. A semantic-based model to represent multimedia big data. In *Proceedings of the 10th international conference on management of digital ecosystems*. 31–38.

[33] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision* 123 (2017), 94–120.

[34] Luca Rossetto, Matthias Baumgartner, Narges Ashena, Florian Ruosch, Romana Pernisch, Lucien Heitz, and Abraham Bernstein. 2021. VideoGraph–towards using knowledge graphs for interactive video retrieval. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II*. Springer, 417–422.

[35] Peter Schallauer, Werner Bailer, Raphaël Troncy, and Florian Kaiser. 2011. Multimedia metadata standards. *Multimedia semantics: Metadata, analysis and interaction* (2011), 129–144.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[37] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20020–20029.

[38] Marian Ursu, Davy Smith, Jonathan Hook, Shauna Concannon, and John Gray. 2020. Authoring Interactive Fictional Stories in Object-Based Media (OBM). In *ACM International Conference on Interactive Media Experiences*. 127–137.

[39] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. 2014. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*. 157–166.

[40] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 322–330.

[41] Mark Williams and John Bell. 2021. The Media Ecology Project: Collaborative DH Synergies to Produce New Research in Visual Culture History. *DHQ: Digital Humanities Quarterly* 15, 1 (2021).

[42] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[43] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment. *arXiv preprint arXiv:2209.06430* (2022).

[44] Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 471–487.