VideoGraph – Towards using Knowledge Graphs for Interactive Video Retrieval

Luca Rossetto[1][0000−0002−5389−9465], Matthias Baumgartner[1][0000−0002−8881−4492], Narges Ashena[1][0000−0002−2820−8159], Florian Ruosch[1][0000−0002−0257−3318], Romana Pernisch[1][0000−0001−8590−1817], Lucien Heitz[1,2][0000−0001−7987−8446], and Abraham Bernstein[1][0000−0002−0128−4602]

[1] Department of Informatics, University of Zurich, Zurich, Switzerland
{lastname}@ifi.uzh.ch
[2] Digital Society Initiative University of Zurich

Abstract. Video is a very expressive medium, able to capture a wide variety of information in different ways. While there have been many advances in the recent past, which enable the annotation of semantic concepts as well as individual objects within video, their larger context has so far not extensively been used for the purpose of retrieval. In this paper, we introduce the first iteration of VideoGraph, a knowledge graph-based video retrieval system. VideoGraph combines information extracted from multiple video modalities with external knowledge bases to produce a semantically enriched representation of the content in a video collection, which can then be retrieved using graph traversal. For the 2021 Video Browser Showdown, we show the first proof-of-concept of such a graph-based video retrieval approach.

Keywords: Interactive Video Retrieval · Knowledge-Graphs · Multi-modal Graphs

1 Introduction

Video is inherently able to capture various information in diverse ways. With increasing advances in machine learning-based content analysis techniques, it becomes increasingly possible to extract and annotate a large amount of this information. While much progress has been made toward said means of information extraction, the integration of such information in the context of multimedia data and for purposes of organization or retrieval of multimedia documents remains understudied. For this edition of the Video Browser Showdown [11], we introduce *VideoGraph*, a Knowledge Graph based video retrieval prototype. Based on similar approaches introduced in *LifeGraph* [10,9] at the Lifelog Search Challenge 2020 [5], VideoGraph uses graph exploration techniques to query a graph composed of information extracted from the challenge dataset [2,14] combined with general knowledge bases [15] which contain general information about the

world. This combination enables querying for richer concepts and situations as those which can be detected directly using currently available methods.

In the remainder of this paper, we first describe the methods used to construct the graph in Section 2 before outlining the methods used for querying it in Section 3. Finally, Section 4 offers some outlook and concluding remarks.

2 VideoGraph Construction

To construct the graph, we make use of information which is directly part of the video dataset [2,14], information which can be extracted from the videos by various means as well as external knowledge bases which provide additional context. The following provides an overview of all these data sources and how they contribute to the graph.

2.1 Wikidata

Wikidata [15], a sister project to the Wikipedia, is *"a free and open knowledge base that can be read and edited by both humans and machines"*.[3] It stores structured data in a graph form which is continuously expanded by a large, international community of volunteers. Since Wikidata can be seen as a general knowledge base without any particular restriction in topics, we use it as a backbone for the construction of the semantic relations between all the concepts extracted from the video dataset. Due to the large diversity in content found within the V3C dataset, we do not use any additional external knowledge bases. This is in contrast to LifeGraph, which also made use of the "Classification of Everyday Living" (COEL) [3].

2.2 Semantic Video Metadata

The videos in the dataset come already with some semantic metadata pulled directly from Vimeo. This includes user-defined tags, Vimeo categories, and textual descriptions. Categories and tags will be added to the graph as resources, therefore linking together videos with the same categories/tags, making their retrieval simpler. However, we will enrich the tags and categories by employing a simple link to Wikidata. This makes it possible, to not only query user defined tags and categories but also include similar concepts as defined within Wikidata.

Even though not all videos include descriptions, we extract entities from the latter where applicable. This adds semantic information where available. Unfortunately, this does not solve the problem for videos which lack descriptions and also do not have many tags or categories attached to them. For these, we have to rely heavily on other forms of information extraction, as described below.

To represent semantic information describing the video contents in the graph, we will rely on existing standards as much as possible [1].

[3] `https://www.wikidata.org/wiki/Wikidata:Main_Page`

2.3 Textual Semantic Information from Video

Textual information in the videos is a rich source to extract *VideoGraph*'s entities from. We use textual information extracted from the V3C1 videos using Optic Character Recognition (OCR) and Automated Speech Recognition (ASR) technologies. We consider two sources for the textual information. First, the speech in the videos and, second, the visible text within the scenes. For the former, we make use of previously extracted, publically available data.[4]

The latter is also based on data already generated for [13]. Neither OCR nor ASR data are however perfect. There are missing values for videos with little to no dialogue or a language different than English. Also, the generated ASR/OCR data is noisy and does not perfectly reflect the speech/text in the videos.

We nevertheless apply entity extraction to both text sources in order to be able to link the contained information to the rest of the graph. In addition to the textual sources, we also use lower-level semantic information which resulted from prior analyses [2] of the dataset.[5]

2.4 Visual Semantic Information from Video

For the extraction of semantic information from the visual component of the videos, we primarily rely on previously generated semantic annotations [13] which were produced by the Google Cloud Vision API.[6] The detectors provided by the API were applied to the representative frame of every video segment in the dataset and for each produced non-localized semantic labels with unique ids. The ids used by the service were already known to Wikidata, so they could be easily mapped to the relevant QIDs. This mapping enables us to link the video segments via their contained semantic concepts not only to each other but also to other concepts, which may not be directly detectable.

2.5 Technical Video Metadata

In order to support an end-to-end retrieval process, the graph does not only need to contain semantic information describing the content of the videos, but also the technical information which is required to interact with the video files themselves, including filenames, codec and frame rate information, shot boundaries, etc. To capture this information, we use the Ontology for Media Resources [6] wherever applicable and make extensions where necessary. Since the graph does not only need to know about every video as a whole but also about every annotated shot, we use the Media Fragments Ontology [8] to represent every video shot from the dataset as a graph node.

[4] `https://github.com/lucaro/V3C1-ASR`
[5] `https://github.com/klschoef/V3C1Analysis`
[6] `https://cloud.google.com/vision`

3 VideoGraph Exploration

Based on the graph described in Section 2, the proposed system supports several means of query construction and subsequent graph exploration methods.

3.1 Query formulation

To retrieve videos, a user can select an arbitrary number of tags of which each corresponds to a node in the graph. Any selection of such tags then needs to be translated into a query against the graph. We follow the approach introduced in [10], where we fill a query template with the selected tags.

In this implementation, we aim to improve the retrieval of videos by adding negation. The user can select tags, which specifically should not be linked to the videos.

Additionally, we provide the user with a raw text input to search specifically within the transcribed audio and detected textual elements in the video. Since these are not tags, we use a different mechanism for querying.

Lastly, in contrast to [10] tags can have different origins. Therefore, we provide an interface, where the user can select the origin of a specific tag. However, this origin is optional.

3.2 Graph exploration

Analogously to the approach described in [10], graph exploration will start at the nodes which have been specified in the query. The graph is then traversed with increasing depth from each start node until either a sufficiently large number of video segments or a maximum depth is reached. The results of the traversals from all start nodes are aggregated and the resulting video segments are scored by their inverse distance to the initial nodes, favoring segments with the shortest distance to the highest number of start nodes.

The full-text components of the query are evaluated independently of the graph since no semantic expansion is necessary there. The results of the two processes are independently sent to the user interface, where their relative weight can be adjusted dynamically.

3.3 Graph extension

Content-based filtering techniques are employed to extend the existing graph. Using correlation-based similarity of video tags [7], additional edges between video nodes with similar tags are introduced. The similarity score for two tags is calculated based on their co-occurrence count in videos.

Furthermore, cosine distances between nodes are leveraged for additional graph extension. Vector cosine-based similarity is used to match video nodes that are alike [7]. The similarity score of videos is based on the assigned video tags they share. The tags create a projection space for finding neighboring nodes. The extended graph will account for neighbors discovered this way by introducing additional edges between the respective nodes.

3.4 User Interaction

Analogously to LifeGraph, the user interface is a modified version of *vitrivr-ng* [4], a browser-based component taken from the content-based multimedia retrieval stack *vitrivr* [12]. vitrivr-ng enables multiple forms of query formulation, of which only the text-based modalities are relevant for our application. Queries on VideoGraph consist of a collection of tags, which can either refer to semantic concepts which have been detected directly or indirectly, as well as free text. The means offered by vitrivr-ng for browsing the retrieved results as well as the UI-based re-ordering and filtering mechanisms are adopted without any major changes.

The UI also offers late-filtering functionality which enables a user to hide a subset of already retrieved results based on Boolean filter criteria which can be applied to metadata associated to the individual results. In order to make optimal use of this functionality, VideoGraph will expose various properties of the individual video shots, both semantic as well as technical, to this filtering mechanism.

4 Conclusion

In this paper, we introduced VideoGraph, our first attempt at representing the complex content of a video collection as a knowledge graph. Guided by insights gained from LifeGraph, which introduced knowledge graph-based exploration of lifelog data, VideoGraph is an initial proof-of-concept with the aim of evaluating the capabilities and limitations of such a graph representation in combination with reasonably simple graph traversal mechanisms for querying. Insights gained from this first evaluation of such a knowledge graph-based video retrieval approach should be able to inform future developments into more complex knowledge representations, such as graph-video co-embeddings.

Acknowledgements

This work was partially funded by the University of Zurich, the Digital Society Initiative, the Swiss Re Institute, and the Swiss National Science Foundation under contract number 200020_184994.

References

1. Arndt, R., Troncy, R., Staab, S., Hardman, L.: Comm: A core ontology for multimediaannotation. In: Handbook on Ontologies (2009)
2. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3c1 dataset: An evaluation of content characteristics. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 334–338 (2019)
3. Bruton, P., Langford, J., Reed, M., Snelling, D.: Classification of everyday living version 1.0 (2019), `https://docs.oasis-open.org/coel/COEL/v1.0/os/COEL-v1.0-os.html`, last updated 23 January 2019

4. Gasser, R., Rossetto, L., Schuldt, H.: Towards an all-purpose content-based multimedia information retrieval system. arXiv preprint arXiv:1902.03878 (2019)
5. Gurrin, C., Le, T.K., Ninh, V.T., Dang-Nguyen, D.T., Jónsson, B.T., Lokoč, J., Hurst, W., Tran, M.T., Schoeffmann, K.: An Introduction to the Third Annual Lifelog Search Challenge, LSC'20. In: ICMR '20, The 2020 International Conference on Multimedia Retrieval. ACM, Dublin, Ireland (2020)
6. Lee, W., Bailer, W., Bürger, T., Champin, P.A., Evain, J.P., Malaisé, V., Michel, T., Sasaki, F., Söderberg, J., Stegmaier, F., et al.: Ontology for media resources 1.0. W3C recommendation **9** (2012)
7. Manjula, R., Chilambuchelvan, A.: Cotent based filtering techniques in recommendation systems using user preferences. International Journal of Innovations in Engineering and Technology **7**(4), 149–154 (2016)
8. Mannens, E., Deursen, D.V., Troncy, R., Pfeiffer, S., Parker, C., Lafon, Y., Jansen, J., Hausenblas, M., Walle, R.: A uri-based approach for addressing fragments of media resources on the web. Multimedia Tools and Applications **59**, 691–715 (2010)
9. Rossetto, L., Baumgartner, M., Ashena, N., Ruosch, F., Pernischova, R., Bernstein, A.: A knowledge graph-based system for retrieval of lifelog data. In: International Semantic Web Conference. pp. 223–228. No. 2721 in Proceedings of the ISWC 2020 Demos and Industry Tracks, CEUR-WS (2020), `http://ceur-ws.org/Vol-2721/paper557.pdf`
10. Rossetto, L., Baumgartner, M., Ashena, N., Ruosch, F., Pernischová, R., Bernstein, A.: Lifegraph: A knowledge graph for lifelogs. In: Proceedings of the Third Annual Workshop on Lifelog Search Challenge. pp. 13–17 (2020)
11. Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., et al.: Interactive video retrieval in the age of deep learning-detailed evaluation of vbs 2019. IEEE Transactions on Multimedia (2020)
12. Rossetto, L., Giangreco, I., Tanase, C., Schuldt, H.: vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 1183–1186 (2016)
13. Rossetto, L., Parian, M.A., Gasser, R., Giangreco, I., Heller, S., Schuldt, H.: Deep learning-based concept detection in vitrivr. In: International Conference on Multimedia Modeling. pp. 616–621. Springer (2019)
14. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3c–a research video collection. In: International Conference on Multimedia Modeling. pp. 349–360. Springer (2019)
15. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)