VideoCLIP 2.0: An Interactive CLIP-Based Video Retrieval System for Novice Users at VBS2024

Thao-Nhu Nguyen[1*], Le Minh Quang[1*], Graham Healy[1], Binh T. Nguyen[2], and Cathal Gurrin[1]

[1] Dublin City University, Ireland
[2] Vietnam National University, Ho Chi Minh University of Science, Vietnam

Abstract. In this paper, we present a revised interactive video retrieval system named VideoCLIP 2.0 developed for the Video Browser Showdown 2024. Building upon the foundation of the previous year's system, VideoCLIP, this upgraded version incorporates several enhancements to support users in solving retrieval tasks. Firstly, the revised system enables search using a variety of modalities, such as rich text, dominant colour, OCR, query-by-image, and now relevance feedback. Additionally, a revised keyframe selection technique has been implemented, as well as a new embedding model to replace the existing CLIP model, aiming to obtain richer visual representations to boost search performance. Lastly, the user interface has been refined to enable quicker inspection and user-friendly navigation, particularly beneficial for novice users.

Keywords: Video Browser Showdown · Interactive Video Retrieval · Embedding Model · Multimodal Retrieval · Video Retrieval System

1 Introduction

In the multimedia content development era, millions of videos have been produced every day which transform how information is shared and consumed. As videos become an increasingly integral part of our digital interactions, the need for efficient and precise video retrieval systems has become undeniable. The annual Video Browser Showdown (VBS) [4] challenge at the MMM conference aims to encourage comparative benchmarking of interactive video retrieval systems in a live, metrics-based evaluation. Similar to previous years' competitions, the participants in VBS are tasked with solving three categories: textual Known-Item Search (t-KIS), visual Known-Item Search (v-KIS), and Ad-hoc Video Search (AVS). While the first two tasks focus on identifying the videos/moments matching the given textual or visual description as fast as possible, the latter's objective is to find as many target moments as possible. The data for this year's challenge consists of a combination of four datasets, including V3C1 (7,475 videos with a duration of 1,000 hours) [2], V3C2 (9,760 videos with a duration of 1,300 hours) [8], Marine Video Kit (MVK) (1,374 videos with a duration of 12.38 hours) [11], and a small set of laparoscopic gynecology videos (LapGynLHE dataset).

[*] Two authors contributed equally to this research.

Traditionally, the systems have usually relied on visual concepts extracted from the video itself, including objects, text, colors, and automatic annotations [1,5]. With the advent of large joint embedding models, these models have been found in many research fields, particularly in vision applications, due to their ability to connect visual content and natural language. Video retrieval is among the applications that have been gaining increasing attention in most retrieval systems [1,9,5,10] in recent years.

Recognising the significance of user-centric design in video retrieval systems, recent research has dived into understanding user behavior, preferences, and expectations. Specifically, we attempt to ease the user experience, especially for novice users who have little to no knowledge about the field. The VideoCLIP 2.0 system, introduced in this paper, represents a refinement of our previous system, VideoCLIP, which performed well at the VBS'23 competition and was ranked overall top search engine at the related IVR4B challenge at the CBMI conference in 2023.

This paper outlines a revised interactive video retrieval system named VideoCLIP 2.0, developed for the Video Browser Showdown 2024. Precisely, we introduce some new functionalities aimed at better facilitating novice users in this year's system. These are the inclusion of a relevance feedback mechanism to support more-like-this search functionality, a revised and improved keyframe selection mechanism to improve the UI, a revised and updated embedding model to support more effective search, and a number of user interface enhancements to better support novice users.

2 An Overview of VideoCLIP 2.0

Our system implements the same architecture as the 2023 system VideoCLIP. Full details can be found in [5]. In summary, VideoCLIP 2.0 builds upon VideoCLIP's underlying engine [6] with the following important modifications. Firstly, the inclusion of a relevant feedback mechanism to facilitate user engagement. Secondly, we use an updated version of the CLIP [7] model, named Open VCLIP [13], to enrich our system's capabilities in supporting multimodal search and retrieval. Thirdly, a new search modality, meta-search, is incorporated, empowering users to filter results with a list of comparative expressions. Lastly, recognising the importance of user experience, we redesign the user interface (UI) to ease the search process by modifying the result presentation, described in Section 2.2.

2.1 Search Modalities

Relevance Feedback Mechanism

Relevance feedback is often used in interactive retrieval systems to facilitate users to provide feedback on the relevance of seen search results, which is then used to refine the search query and hopefully further improve the results. In VideoCLIP 2.0 we employ relevance feedback to facilitate the selection of any

keyframe. This process involves appending the metadata linked to the selected keyframe to the existing user query, ultimately forming a new ranked list. This iterative approach enables users to actively shape and fine-tune their search experience based on the perceived relevance of visual content.

Improved Keyframe Selection

Given the interactive use of the VideoCLIP 2.0 system, it becomes increasingly important for the system to select representative keyframes to represent groups of sequential video shots. For this year's system, we employ the CLIP model to assist in the selection of query-relayed keyframes for display on the screen during a user's search session. The similarity between the query and keyframe embeddings is computed for a series of query-relevant shots and the frames with the highest similarity scores are selected for display to the user. In this way, we can optimise the use of screen real estate by not showing every shot keyframe from a highly ranked video or video segment.

Enhanced Embedding Models

Due to its ability to link between visual concepts and natural language, the CLIP model [7] is widely used in many research fields, including image classification, image similarity, and image captioning. The CLIP has proven itself to be the state-of-the-art embedding model and utilised by various teams during the challenge, including Vibro [3] and our own VideoCLIP [5], which were the winners of the VBS2022 and VBS2023, respectively.

Nevertheless, the aforementioned CLIP model has its own drawbacks. That is, while CLIP can effectively identify a new action or object appearing in the testing images, it cannot do the same for unforeseen events in videos, which leads to its unsuitability when being used in our system that uses video data as the input. Thus, for our previous work in [5], a technique called "frame-based retrieval" has been applied. To be more specific, for each of the videos being considered, a frame is picked and treated as its representative. This approach is proven to have many disadvantages. One of them is that the chosen frame may not cover every information of the rooted video. Also, which frame from the video is picked can have a notable effect on the overall results, leading to the inconsistency of the system in general. Therefore, it is necessary to find a model that possesses the strengths of the CLIP model, while can also adapt better to problems related to video data.

In our VideoCLIP 2.0 system, Open-VCLIP [12], a modified and improved version of the CLIP model introduced by Weng et al., is considered. The latest version of this model, namely Open-VCLIP++ [13] was released by Wu et al. in October 2023. In this model, a method named Interpolated Weight Optimization is introduced to leverage the weight interpolation during training and testing phrases. In the mentioned papers, the authors also show the significant strength of the model compared to the other methods. Thus, the Open-VCLIP is expected to be a suitable alternative option to the already used CLIP model in our newly proposed system.

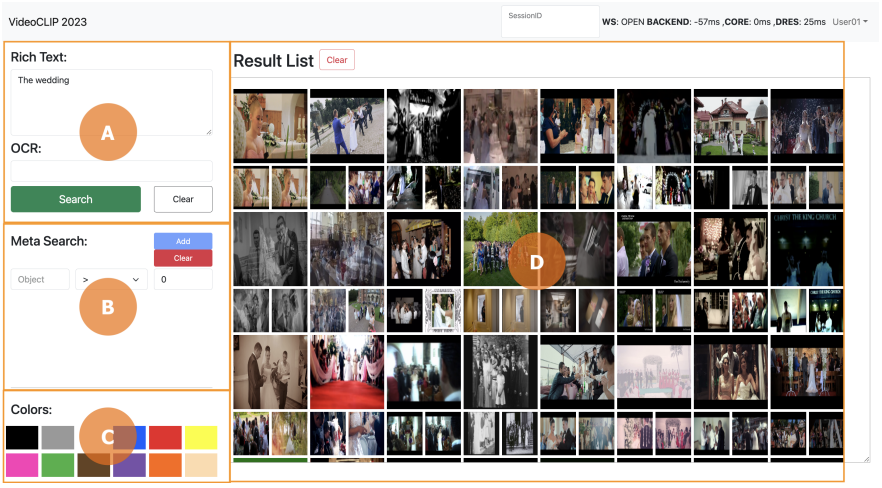

Fig. 1: The prototype of VIDEOCLIP 2.0 User Interface

2.2 User Interface Revision

After participating in VBS2023, we recognise that the time and steps taken for search and submission are critical to the system's overall performance. To enhance this aspect, we have updated the UI with minor revisions intending to enhance both the speed and user-friendliness of inspecting ranked lists.

While maintaining the foundational layout from the previous system, our focus is on simplifying the UI to provide users, particularly novices, with not only a better experience but also an improvement in search performance. Specifically, users can now input and formulate queries on the left side of the interface, while the results are conveniently displayed on the right side. By prioritizing and keeping only the most crucial buttons, we aim to enhance user clarity and expedite their interaction with the system.

3 Conclusion

In this paper, we present the latest version of our video interactive retrieval system, VIDEOCLIP 2.0, a CLIP-based system. We have updated the user interface, and the search result visualisation, along with upgrading the back end with the latest CLIP models. In addition, a meta-search is added to alleviate possible weaknesses of the CLIP model.

Acknowledgments

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 18/CRT/6223, and 13/RC/2106_P2 at the ADAPT SFI Research Centre at DCU. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme.

References

1. G. Amato, P. Bolettieri, F. Carrara, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, and C. Vairo. Visione at video browser showdown 2023. In D.-T. Dang-Nguyen, C. Gurrin, M. Larson, A. F. Smeaton, S. Rudinac, M.-S. Dao, C. Trattner, and P. Chen, editors, *MultiMedia Modeling*, pages 615–621, Cham, 2023. Springer International Publishing.
2. F. Berns, L. Rossetto, K. Schoeffmann, C. Beecks, and G. Awad. V3C1 Dataset: An Evaluation of Content Characteristics. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, page 334–338, New York, NY, USA, 2019. Association for Computing Machinery.
3. N. Hezel, K. Schall, K. Jung, and K. U. Barthel. Efficient Search and Browsing of Large-Scale Video Collections with Vibro. In B. Þór Jónsson, C. Gurrin, M.-T. Tran, D.-T. Dang-Nguyen, A. M.-C. Hu, B. Huynh Thi Thanh, and B. Huet, editors, *MultiMedia Modeling*, Lecture Notes in Computer Science, page 487–492, Cham, 2022. Springer International Publishing.
4. J. Lokoč, P. Veselý, F. Mejzlík, G. Kovalčík, T. Souček, L. Rossetto, K. Schoeffmann, W. Bailer, C. Gurrin, L. Sauter, J. Song, S. Vrochidis, J. Wu, and B. t. Jónsson. Is the reign of interactive search eternal? findings from the video browser showdown 2020. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(3), jul 2021.
5. T.-N. Nguyen, B. Puangthamawathanakun, A. Caputo, G. Healy, B. T. Nguyen, C. Arpnikanondt, and C. Gurrin. Videoclip: An interactive clip-based video retrieval system at vbs2023. In D.-T. Dang-Nguyen, C. Gurrin, M. Larson, A. F. Smeaton, S. Rudinac, M.-S. Dao, C. Trattner, and P. Chen, editors, *MultiMedia Modeling*, pages 671–677, Cham, 2023. Springer International Publishing.
6. T.-N. Nguyen, B. Puangthamawathanakun, G. Healy, B. T. Nguyen, C. Gurrin, and A. Caputo. Videofall - a hierarchical search engine for vbs2022. In B. Þór Jónsson, C. Gurrin, M.-T. Tran, D.-T. Dang-Nguyen, A. M.-C. Hu, B. Huynh Thi Thanh, and B. Huet, editors, *MultiMedia Modeling*, pages 518–523, Cham, 2022. Springer International Publishing.
7. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.
8. L. Rossetto, K. Schoeffmann, and A. Bernstein. Insights on the V3C2 Dataset. *CoRR*, abs/2105.01475, 2021.
9. L. Sauter, R. Gasser, S. Heller, L. Rossetto, C. Saladin, F. Spiess, and H. Schuldt. Exploring effective interactive text-based video search in vitrivr.
10. K. Schoeffmann, D. Stefanics, and A. Leibetseder. divexplore at the video browser showdown 2023. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*, pages 684–689. Springer, 2023.
11. Q.-T. Truong, T.-A. Vu, T.-S. Ha, J. Lokoč, Y. H. W. Tim, A. Joneja, and S.-K. Yeung. Marine video kit: A new marine video dataset for content-based analysis and retrieval. In *MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023*. Springer, 2023.

12. Z. Weng, X. Yang, A. Li, Z. Wu, and Y.-G. Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *ICML*, 2023.
13. Z. Wu, Z. Weng, W. Peng, X. Yang, A. Li, L. S. Davis, and Y.-G. Jiang. Building an open-vocabulary video clip model with better architectures, optimization and data, 2023.