VidBasys: A User-friendly Interactive Video Retrieval System for Novice Users in IVR4B

Thao-Nhu Nguyen*
School of Computing
Dublin City University
Dublin, Ireland
thaonhu.nguyen24@mail.dcu.ie

Quang-Linh Tran*
ADAPT Centre, School of Computing
Dublin City University
Dublin, Ireland
linh.tran3@mail.dcu.ie

Hoang-Bao Le
ADAPT Centre, School of Computing
Dublin City University
Dublin, Ireland
bao.le2@mail.dcu.ie

Binh T. Nguyen
University of Science
Vietnam National University
Ho Chi Minh City, Vietnam
ngtbinh@hcmus.edu.vn

Liting Zhou
ADAPT Centre, School of Computing
Dublin City University
Dublin, Ireland
liting.zhou@dcu.ie

Gareth J. F. Jones
Cathal Gurrin
ADAPT Centre, School of Computing
Dublin City University
Dublin, Ireland
{gareth.jones, cathal.gurrin}@dcu.ie

***Abstract*—In this paper, we present the VidBasys interactive video retrieval system for novice users, an upgraded version of VideoCLIP 2.0 that participated in the Video Browser Showdown 2024. While the novel user interface is designed in a more user-friendly way for newbies to accommodate the target of the Interactive Video Retrieval for Beginner (IVR4B), the core search engine is enhanced with the advance of the recent CLIP model to bridge the gap in semantics between image and text. This version is designed to focus on novice users with a simple, easy-to-use but effective user interface. The system supports free-text search to enhance the user experience and minimise the number of actions required for filtering. The new user interface supports simple search and filters with clearly designed free-text search boxes. In addition, the retrieved results are displayed in an optimised layout to maximise image display space and minimise user interactions. The improvements are expected to support novice users in accurately retrieving the desired videos.**

***Index Terms*—video retrieval, multi-media analysis, user interface**

I. Introduction

In the digital era dominated by the development of social media platforms such as YouTube and TikTok, video content production has made significant progress. The large volume of visual content data brings significant challenges in managing and retrieving. To address this need, there is a growing demand for large-scale content-based retrieval systems capable of efficiently organising and navigating through extensive video collections. These systems play a vital role in various fields, from entertainment to education and healthcare, where they allow recommendations for personalised content, facilitate access to instructional resources, and aid in medical imaging analysis [1].

Interactive Video Retrieval (IVR) is a field within multimedia content analysis that focuses on enabling users to efficiently search and retrieve relevant video content through interactive interfaces. Recent advances in artificial intelligence (AI) and computer vision (CV) have significantly improved the capabilities of video retrieval systems, allowing for more accurate and efficient content analysis by multi-modal models such as CLIP [2], and BLIP [3]. However, despite these advancements, there remains a pressing demand for user-friendly retrieval systems, especially for novice users who have little to no knowledge of retrieval in general, and may lack technical expertise in navigating complex search interfaces in particular. Benchmark challenges such as the Video Browser Showdown (VBS) [4], [5] and the Lifelog Search Challenge (LSC) [6], [7] serve as critical platforms for evaluating the performance of visual-content retrieval systems. Interactive Video Retrieval for Beginners (IVR4B), is one of these challenges, it utilises the video dataset from VBS but focuses specifically on beginners in the field of video retrieval. This challenge utilises the dataset from the VBS2024 competition with four datasets, with a total of 18,684 videos, including V3C1 (7,475 videos) [8], V3C2 (9,760 videos) [9], Marine Video Kit (MVK) (1,374 videos) [10], and a small set of laparoscopic gynaecology videos (LapGynLHE dataset) (75 videos). These large and diverse datasets are a big challenge to retrieve accurately, especially for novice users.

Inherited from the VideoCLIP 2.0 [11] system that participated in the VBS2024 competition, we built upon that platform with some new adjustment functions for novice users such as visual similarity. A big improvement in this VidBasys system is the new user interface (UI). The interface is designed with intuitive elements designed for the user's convenience. At its core is a free-text search feature, thoughtfully arranged to eliminate the need for intricate cognitive processes. Retrieved keyframes are presented in a grid format to offer a comprehensive view for users to explore the results. Furthermore, includ-

* Thao-Nhu Nguyen and Quang-Linh Tran contributed equally to this work.

ing the relevant filter box for Optical Character Recognition (OCR), enhances search precision and efficiency. In this new interface, VidBasys introduces a visual similarity search page, empowering users to explore content through image queries. These new features are expected to help newbies in video retrieval find the desired videos with the least effort and the most accuracy.

II. Related Works

In the development of artificial intelligence, machine learning systems are designed to apply multiple approaches and require more human impact on the model process. Following this idea, Video Browser Showdown [4], [5] (VBS) and Lifelog Search Challenge [6], [7] (LSC) have been founded, and last year was Interactive Video Retrieval for Beginners (IVR4B). Since the first established days, these competitions have attracted many teams and become an annual playground for them to improve and gain better results. Many systems have been introduced such as LifeSeeker [12], MyEachtra [13] and MemoriEase [14], [15] in LSC, and diveXplore [16], [17], Exquisitor [18] and vitrivr [19], [20] in VBS.

Inspired by the concept of VBS and LSC, IVR4B was organised for the first time last year, and the main goal of this competition is to build an easy-to-use video retrieval system for users without knowing the underground architecture. The first system, diveXB [21], which achieved the Best KIS Visual System, has a three-part architecture: backend, middleware and frontend. In addition to using OpenCLIP [22] instead of CLIP [23] and other back-end improvements, diveXB is implemented as an Angular web GUI that makes it well suited for beginners. Nominated as Best KIS Textual System, Exquisitor [18] has introduced UI changes to accelerate the user's experience and solve their task better. There is a Screen Utilisation feature, which maximises the usage of available screen space. The Temporal Queries are designed around building additional relevant feedback classifiers, that focus on the different events described in the task, and then merging the results of the classifiers. Finally, the Video Summary features timelines for the shots in the video, and a video player playing the selected shot. As the runner-up of VBS 2023, Amato et al. [24] built the VISIONE version for newbies. This system is inspired by well-known search engines such as Google and Bing to provide a simple and user-friendly interface for novice users. VISIONE supports users with three search functionalities: free text search, spatial colour and object search, and similarity search. The authors implemented three similarity searches: based on GEM features [25], based on ALADIN [26] features for video keyframes, and based on CLIP2Video [27] features for video clips. Sauter et al. [28] redesigned the vitrivr [20] into the minimal one named vitrivr$_{\text{min}}$. This system offers three means of querying for specific information contained in the video Text on Screen, Speech, and Scene description. For the Text-on-Screen querying, authors apply HyText [29] an optical character recognition method optimised for video. Moreover, they also utilise the 'whisper' [30] transcription and the CLIP embedding [2] for the Speech and Scene description querying respectively.

In addition to the four systems mentioned above, we also participated in IVR4B last year with an interactive video retrieval system - VideoClip [31] and gained the Best Overall System award. Our system workflow is inherited from VideoFall [32] and added two different search modalities for supporting non-expert users: a free-text search, and a query-by-example search. In addition, the user interface has been improved with better visuals and additional features to improve user experience. Inheriting the idea from the last system and incorporating new features from VideoCLIP 2.0 [11] at VBS2024, VidBasys is released with both a new user interface and adjustments in the underlining architecture. The new user interface with a simple but efficient layout helps novice users to search and locate desired videos effectively. The improvements are expected to help newbies accurately retrieve the desired videos.

III. System Overview

In this section, we present the architecture of the interactive video retrieval system and outline several enhancements that are applied to this system. Figure 1 depicts an overview of the system. We use the extracted features such as OCR and colour to filter the retrieved results. The CLIP encoders extract the embeddings of keyframe images and query before calculating the cosine similarity between them. The top matches with the highest cosine similarity are then filtered by OCR and colour before returning to the user interface. More details about each component are described in this section.

A. CLIP Embedding

In this system, we use the Open-VCLIP model [33] to perform the embedding-based retrieval thanks to its significant implications in bridging the semantic gap between images and text. Contrastive Language-Image Pretraining (CLIP) [2] is a pioneering embedding model designed to understand and represent the relationship between visual and textual data. It projects the visual and textual data to a mutual latent representation. From that, we can compare the similarity of them to perform the retrieval. Using this CLIP family, we employ Open-VCLIP as the main embedding model. Open-VCLIP incorporates enhancements and modifications to the original CLIP architecture, resulting in superior performance, particularly notable in action recognition tasks on video datasets. Open-VCLIP is expected to serve as a better alternative to the established CLIP model in our proposed system.

B. Feature Extraction

Filtering plays an important role in narrowing down the search space and providing more accurate results. We extracted two features that can serve as filtering for the retrieved results, including Optical Character Recognition (OCR) and color detection. Users can type the OCR filtering and choose the colour of the video in the user interface. These filters are then applied in the search stage to find keyframes that only satisfy

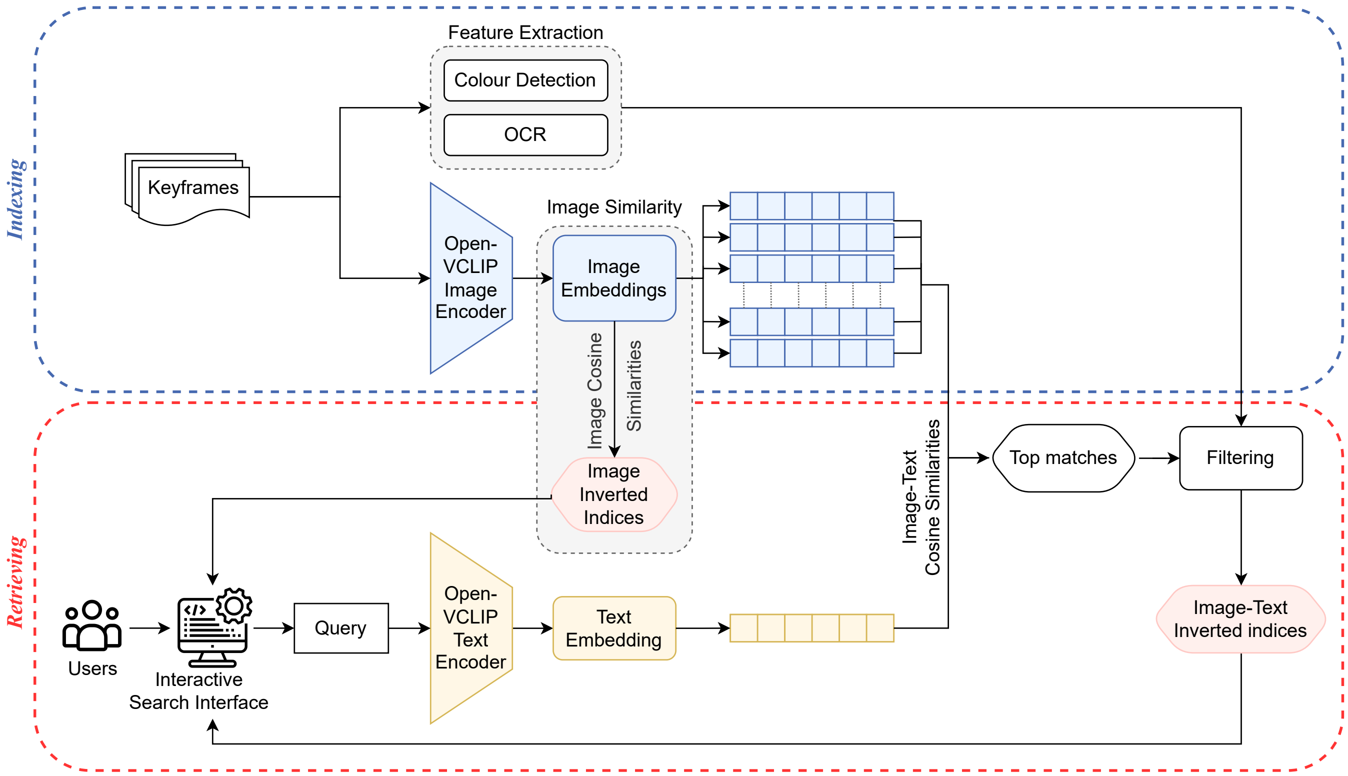


Fig. 1. An overview of the VidBasys workflow, which consists of two stages: Indexing and Retrieving. In the Indexing stage, input keyframes undergo feature extraction for color and text, and are converted into image embeddings. During the Retrieval stage, users submit queries via a search interface, which are converted into text embeddings. Subsequently, the cosine similarities between image and text embeddings are calculated to identify and rank relevant results. Additionally, the image similarity is done by comparing the features of the chosen image with the image embedding collection.

these filters. Any results in the top matches that do not fit the filters are removed.

C. Visual Similarity

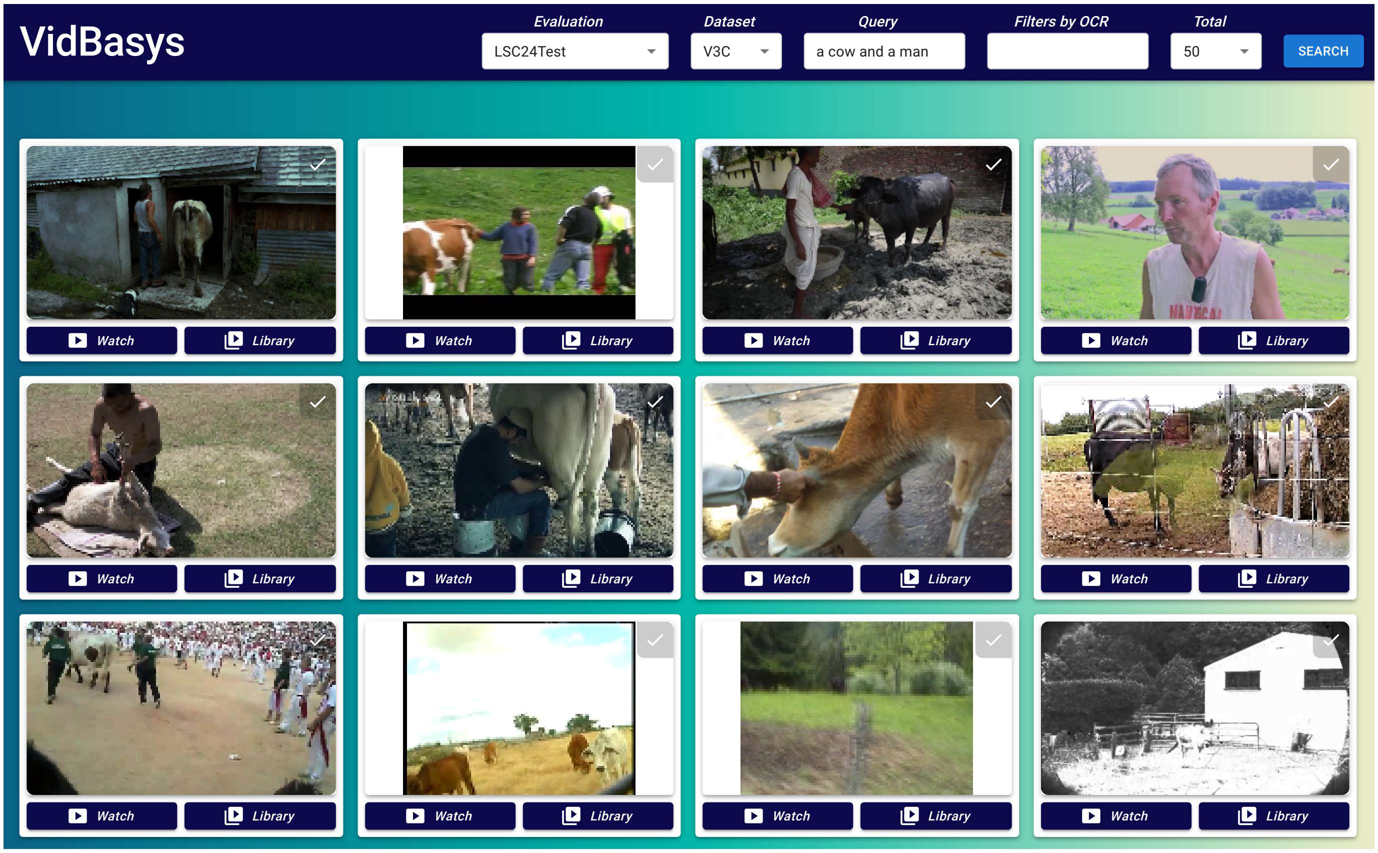
In order to enhance the precision of searching a particular keyframe, the system facilitates visual similarity search. The approach is embedding-based retrieval from image input. We use the CLIP embedding model to derive the embedding of the input keyframe. The embedding of this input frame is subsequently calculated as cosine similarity with embeddings of other frames, and the most relevant matches undergo filtering before being presented to users.

IV. User Interface

In this section, we outline a new user interface of the system, designed to simplify the process of retrieving videos for inexperienced users. This interface is a big change from the previous interface for VideoCLIP2 [11] systems. Instead of representing keyframes in triplicate, we simplify the representation with a single keyframe. Figure 2 illustrates the main search page of the VidBasys system with several notable features, including free text search, visual similarity search (Library button) and display options (Watch buttons). Firstly, the system offers a free-text search feature by the query box for the main query. The query box is accompanied by clear descriptions, minimising the cognitive load required for users to conduct their searches. Secondly, all retrieved keyframes are presented in a grid format to maximise space utilisation, enabling users to view as many videos as possible. Essential actions applicable to the retrieved keyframes, such as submission, full video viewing, and finding similar keyframes, are depicted with relevant icons in the keyframes. Thirdly, a filter is provided next to the search box, allowing users to refine their searches through Optical Character Recognition (OCR). Finally, a new page for visual similarity search is constructed to support users in finding videos by keyframe.

The user interface is divided into two primary sections: the search bar located at the top and the result display located at the bottom. In the upper left corner of the interface, the system's name is prominently displayed within the search bar. Adjacent to it, there is a drop-down filter offering three options for users to choose the video dataset they want to search within. The main query box is centrally placed in the UI, which is the largest free-text search box, enabling users to enter any text in any format to initiate a search. Additionally, situated at the top right of the search bar is the OCR filter box, which allows users to input text to find keyframes containing specific text. The OCR filter is an exact match filter that ensures that only text that precisely matches the user input is retained.

The result-display section of the UI appears below the search bar. It showcases retrieved results in a grid format,



Evaluation: LSC24Test
Dataset: V3C
Query: a cow and a man
Filters by OCR:
Total: 50

Fig. 2. The User Interface of VidBasys system. The screen displayed the result of an example query “a cow and a man” in the V3C dataset.

typically comprising around 3 rows and 4 columns, allowing users to view up to 12 keyframes per page. Each keyframe is equipped with various action icons for user interaction. Specifically, users can submit the keyframe, view the full video, or explore visually similar keyframes, with these icons conveniently positioned at the top right corner and bottom of each keyframe, respectively.

V. Conclusion

In this paper, we introduce an updated version of VideoCLIP 2.0 system that participated in the Video Browser Showdown 2024. This offers an easy-to-use solution for novice users seeking an intuitive and efficient video-retrieving experience. By incorporating a user-friendly interface and enhanced free text search capability, we have reduced the cognitive load and streamlined the video retrieval process for beginners. Leveraging recent advancements in the CLIP model, we have moved close to bridging the semantic gap between image and text, enhancing the accuracy of video retrieval. The modifications implemented prioritise simplicity, aligning with the goals of the IVR4B workshop. In general, these enhancements are expected to empower novice users, enabling them to find the videos they are looking for effortlessly with precision and ease.

Acknowledgment

This research was carried out 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 Center at Dublin City University. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

[1] Z. Li, X. Zhang, H. Müller, and S. Zhang, “Large-scale retrieval for medical image analytics: A comprehensive review,” *Medical image analysis*, vol. 43, pp. 66–84, 2018.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[3] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[4] J. Lokoč, S. Andreadis, W. Bailer, A. Duane, C. Gurrin, Z. Ma, N. Messina, T.-N. Nguyen, L. Peška, L. Rossetto, L. Sauter, K. Schall, K. Schoeffmann, O. S. Khan, F. Spiess, L. Vadicamo, and S. Vrochidis, “Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th vbs,” *Multimedia Syst.*, vol. 29, no. 6, p. 3481–3504, aug 2023. [Online]. Available: https://doi.org/10.1007/s00530-023-01143-5

[5] S. Heller, V. Gsteiger, W. Bailer, C. Gurrin, B. Þ. Jónsson, J. Lokoč, A. Leibetseder, F. Mejzlík, L. Peška, L. Rossetto *et al.*, "Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 1–18, 2022.

[6] C. Gurrin, B. Þ. Jónsson, D. T. D. Nguyen, G. Healy, J. Lokoc, L. Zhou, L. Rossetto, M.-T. Tran, W. Hürst, W. Bailer *et al.*, "Introduction to the sixth annual lifelog search challenge, lsc'23," in *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023, pp. 678–679.

[7] C. Gurrin, L. Zhou, G. Healy, B. Þór Jónsson, D.-T. Dang-Nguyen, J. Lokoć, M.-T. Tran, W. Hürst, L. Rossetto, and K. Schöffmann, "Introduction to the fifth annual lifelog search challenge, lsc'22," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 685–687.

[8] 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*, 2019, pp. 334–338.

[9] L. Rossetto, K. Schoeffmann, and A. Bernstein, "Insights on the v3c2 dataset," *arXiv preprint arXiv:2105.01475*, 2021.

[10] Q.-T. Truong, T.-A. Vu, T.-S. Ha, J. Lokoč, Y.-H. Wong, A. Joneja, and S.-K. Yeung, "Marine video kit: a new marine video dataset for content-based analysis and retrieval," in *International Conference on Multimedia Modeling*. Springer, 2023, pp. 539–550.

[11] T.-N. Nguyen, L. M. Quang, G. Healy, B. T. Nguyen, and C. Gurrin, "Videoclip 2.0: An interactive clip-based video retrieval system for novice users at vbs2024," in *MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 394–399.

[12] T.-N. Nguyen, T.-K. Le, V.-T. Ninh, C. Gurrin, M.-T. Tran, T. B. Nguyen, G. Healy, A. Caputo, and S. Smyth, "E-lifeseeker: An interactive lifelog search engine for lsc'23," in *Proceedings of the 6th Annual ACM Lifelog Search Challenge*, 2023, pp. 13–17.

[13] L. D. Tran, B. Nguyen, L. Zhou, and C. Gurrin, "Myeachtra: Event-based interactive lifelog retrieval system for lsc'23," in *Proceedings of the 6th Annual ACM Lifelog Search Challenge*, 2023, pp. 24–29.

[14] Q.-L. Tran, L.-D. Tran, B. Nguyen, and C. Gurrin, "Memoriease: An interactive lifelog retrieval system for lsc'23," in *Proceedings of the 6th Annual ACM Lifelog Search Challenge*, 2023, pp. 30–35.

[15] Q.-L. Tran, B. Nguyen, G. Jones, and C. Gurrin, "Memoriease at the ntcir-17 lifelog-5 task," 2024.

[16] K. Schoeffmann and S. Nasirihaghighi, "Divexplore at the video browser showdown 2024," in *MultiMedia Modeling*, S. Rudinac, A. Hanjalic, C. Liem, M. Worring, B. Þ. Jónsson, B. Liu, and Y. Yamakata, Eds. Cham: Springer Nature Switzerland, 2024, pp. 372–379.

[17] 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*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 684–689. [Online]. Available: https://doi.org/10.1007/978-3-031-27077-2_59

[18] O. S. Khan and B. o. Jónsson, "User relevance feedback and novices: Anecdotes from exquisitor's participation in interactive retrieval competitions," in *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, ser. CBMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 173–177. [Online]. Available: https://doi.org/10.1145/3617233.3617275

[19] F. Spiess, R. Gasser, S. Heller, M. Parian-Scherb, L. Rossetto, L. Sauter, and H. Schuldt, "Multi-modal video retrieval in virtual reality with vitrivr-vr," in *MultiMedia Modeling*. Cham: Springer International Publishing, 2022, pp. 499–504.

[20] L. Sauter, R. Gasser, S. Heller, L. Rossetto, C. Saladin, F. Spiess, and H. Schuldt, "Exploring effective interactive text-based video search in vitrivr," in *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 646–651. [Online]. Available: https://doi.org/10.1007/978-3-031-27077-2_53

[21] K. Schoeffmann, "divexb: An interactive video retrieval system for beginners," in *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, ser. CBMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 153–157. [Online]. Available: https://doi.org/10.1145/3617233.3617258

[22] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," 2022.

[23] 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," 2021.

[24] G. Amato, P. Bolettieri, F. Carrara, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, and C. Vairo, "Visione for newbies: an easier-to-use video retrieval system," in *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, ser. CBMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 158–162. [Online]. Available: https://doi.org/10.1145/3617233.3617261

[25] J. Revaud, J. Almazan, R. S. de Rezende, and C. R. de Souza, "Learning with average precision: Training image retrieval with a listwise loss," 2019.

[26] N. Messina, M. Stefanini, M. Cornia, L. Baraldi, F. Falchi, G. Amato, and R. Cucchiara, "Aladin: Distilling fine-grained alignment scores for efficient image-text matching and retrieval," 2022.

[27] H. Fang, P. Xiong, L. Xu, and Y. Chen, "Clip2video: Mastering video-text retrieval via image clip," 2021.

[28] L. Sauter, H. Schuldt, R. Waltenspül, and L. Rossetto, "Novice-friendly text-based video search with vitrivr," in *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, ser. CBMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 163–167. [Online]. Available: https://doi.org/10.1145/3617233.3617262

[29] A. Theus, L. Rossetto, and A. Bernstein, "Hytext – a scene-text extraction method for video retrieval," in *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 182–193. [Online]. Available: https://doi.org/10.1007/978-3-030-98355-0_16

[30] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[31] T.-N. Nguyen, B. Puangthamawathanakun, C. Arpnikanondt, C. Gurrin, A. Caputo, and G. Healy, "Efficient search with an interactive video retrieval system for novice users in ivr4b," in *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, ser. CBMI '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 168–172. [Online]. Available: https://doi.org/10.1145/3617233.3617273

[32] T.-N. Nguyen, B. Puangthamawathanakun, G. Healy, B. T. Nguyen, C. Gurrin, and A. Caputo, "Videofall-a hierarchical search engine for vbs2022," in *International Conference on Multimedia Modeling*. Springer, 2022, pp. 518–523.

[33] 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 *International Conference on Machine Learning*. PMLR, 2023, pp. 36 978–36 989.