VISIONE for newbies: an easier-to-use video retrieval system

Giuseppe Amato
giuseppe.amato@isti.cnr.it
ISTI-CNR
Pisa, Italy

Paolo Bolettieri
paolo.bolettieri@isti.cnr.it
ISTI-CNR
Pisa, Italy

Fabio Carrara
fabio.carrara@isti.cnr.it
ISTI-CNR
Pisa, Italy

Fabrizio Falchi
fabrizio.falchi@isti.cnr.it
ISTI-CNR
Pisa, Italy

Claudio Gennaro
claudio.gennaro@isti.cnr.it
ISTI-CNR
Pisa, Italy

Nicola Messina
nicola.messina@isti.cnr.it
ISTI-CNR
Pisa, Italy

Lucia Vadicamo*
lucia.vadicamo@isti.cnr.it
ISTI-CNR
Pisa, Italy

Claudio Vairo
claudio.vairo@isti.cnr.it
ISTI-CNR
Pisa, Italy

ABSTRACT

This paper presents a revised version of the VISIONE video retrieval system, which offers a wide range of search functionalities, including free text search, spatial color and object search, visual and semantic similarity search, and temporal search. The system is designed to ensure scalability using advanced indexing techniques and effectiveness using cutting-edge Artificial Intelligence technology for visual content analysis. VISIONE was the runner-up in the 2023 Video Browser Showdown competition, demonstrating its comprehensive video retrieval capabilities. In this paper, we detail the improvements made to the search and browsing interface to enhance its usability for non-expert users. A demonstration video of our system with the restyled interface, showcasing its capabilities on over 2,300 hours of diverse video content, is available online at https://youtu.be/srD3TCUkMSg.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Users and interactive retrieval**; **Retrieval models and ranking**; *Search engine architectures and scalability*; **Multimedia and multimodal retrieval**; **Video search**; • **Human-centered computing → Interactive systems and tools**.

KEYWORDS

multimedia retrieval, video search, cross-modal search, interactive system, video retrieval, user interface

ACM Reference Format:
Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2023. VISIONE for newbies: an easier-to-use video retrieval system. In *20th International Conference on Content-based Multimedia Indexing (CBMI 2023), September 20–22, 2023, Orléans, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3617233.3617261

1 INTRODUCTION

The explosion of visual content production that we are witnessing in the era of digital cameras and social media platforms has created a significant challenge for both the management of and retrieval in visual archives. This is particularly true for user-generated content, where the visual data is often poorly annotated or not annotated at all. This has also led to a great demand for content-based retrieval systems capable of handling increasingly large multimedia data collections. These systems can have significant implications in various fields, such as journalism, surveillance, and entertainment.

Recent advances in artificial intelligence-based technologies have enabled the development of effective approaches for analyzing and understanding multimedia content. However, there is still a strong demand for user-friendly tools that can integrate these technologies for large-scale multimedia searching in an interactive way. To foster the development of large-scale interactive video retrieval systems, benchmarking competitions such as the Video Browser Showdown (VBS) [14, 16, 18, 19, 27, 28] and the Lifelog Search Challenge [12, 32] are organized annually to evaluate different systems' performance in live video search sessions. These competitions provide a platform for researchers and developers to showcase their work and assess their systems' performance against others. The growing research interest in multimedia retrieval and analysis tasks is evidenced by the increasing participation in these competitions [15, 17, 20, 22, 23, 29–31], highlighting the need for more effective and user-friendly tools to handle the growing amount of multimedia data.

VISIONE is an interactive large-scale video search system that leverages state-of-the-art Artificial Intelligence (AI) techniques for visual content analysis and advanced indexing techniques to ensure scalability. It integrates multiple search functionalities, including free text search, spatial color and object search, visual and semantic

[*] Corresponding author

CBMI 2023, September 20–22, 2023, Orléans, France

ACM ISBN 979-8-4007-0912-8/23/09…$15.00
https://doi.org/10.1145/3617233.3617261

similarity search, and temporal search. The first version of the system was developed in 2019 [1, 2], and after that, it has been updated almost every year [3, 5, 6] in conjunction with participation in the VBS competition.

At the last VBS competition held in January 2023, VISIONE achieved remarkable success in numerous tasks and ranked second in the overall leaderboard. Although its excellent performance in the last VBS demonstrates how VISIONE effectively integrates various search features and thus provides a comprehensive solution for video retrieval, its user interface may not be easy for novice users. With numerous search options available, including the ability to combine them, the interface could be overwhelming and confusing for users unfamiliar with the system.

This paper presents a new iteration of the system, which represents the fifth version since 2019. This version focuses on enhancing the user experience for novice users, accomplished through a redesigned user interface. To evaluate the usability of the new system and assess the effectiveness of our redesign in facilitating usage for non-expert users, we plan to participate in the Interactive Video Retrieval for Beginners (IVR4B) special session at CMBI 2023, employing both the previous system version utilized in VBS 2023 and the version described in this paper.

2 SYSTEM OVERVIEW

VISIONE offers various search functionalities to enable users to search for specific video segments. It includes free text search, spatial color and object search, and similarity search. These search functionalities can be used alone or in combination with a temporal search.

To support free text search, VISIONE uses cross-modal features extracted using three pre-trained models: CLIP2Video [10], CLIP [24] trained on the LAION dataset[1] and ALADIN [21]. Features extracted using CLIP2Video and ALADIN are also employed for similarity search in combination with GEM features [26]. Moreover to facilitate issuing a textual query by dictating it to the system instead of typing it, we integrated a speech-to-text functionality based on the Whisper model [25].

We employ three models for object detection: VfNet [34], Mask R-CNN [13], and a Faster R-CNN [11], trained on COCO, LVIS, and Open Images V4 datasets, respectively. Color annotation is performed using two chip-based color naming techniques [8, 33].

All the extracted features and descriptors, with the exception of the CLIP-based ones, are stored and searched using a full-text search engine. Specifically, we employ Apache Lucene[2]. The CLIP-based features are stored in a second index and searched using the FAISS library[3]. We use two separate indexes due to the different functionalities and implementations required. We originally decided to employ Lucene because its disk-based architecture allows scalability to billions of documents without the need for large main memory requirements, unlike in-memory indexes such as FAISS. Typically, Lucene is utilized for text-based searches in large unstructured text document collections. However, it can also be employed for data encoded as text, such as quantized colors and object classes, including their respective 2D coordinates within the frame, as in our case [2]. To index feature vectors obtained from deep neural networks, we developed a set of techniques known as Surrogate Text Representations (STRs) [2, 7, 9]. These techniques transform dense features into sparse term-frequency vectors derived from a synthetic codebook while preserving the dot product between the resulting textual representation and the original dense feature to the best possible extent. Although STRs have been proven effective for many features, such as GEM and ALADIN, we encountered significant issues when applying this encoding technique to CLIP-based features. This is due to a different distribution of cosine similarity between the query text and the top nearest neighboring images [4]. We are currently working on developing a new STR approach that can better handle CLIP features. Meanwhile, we rely on FAISS for searching through CLIP-based features for convenience.

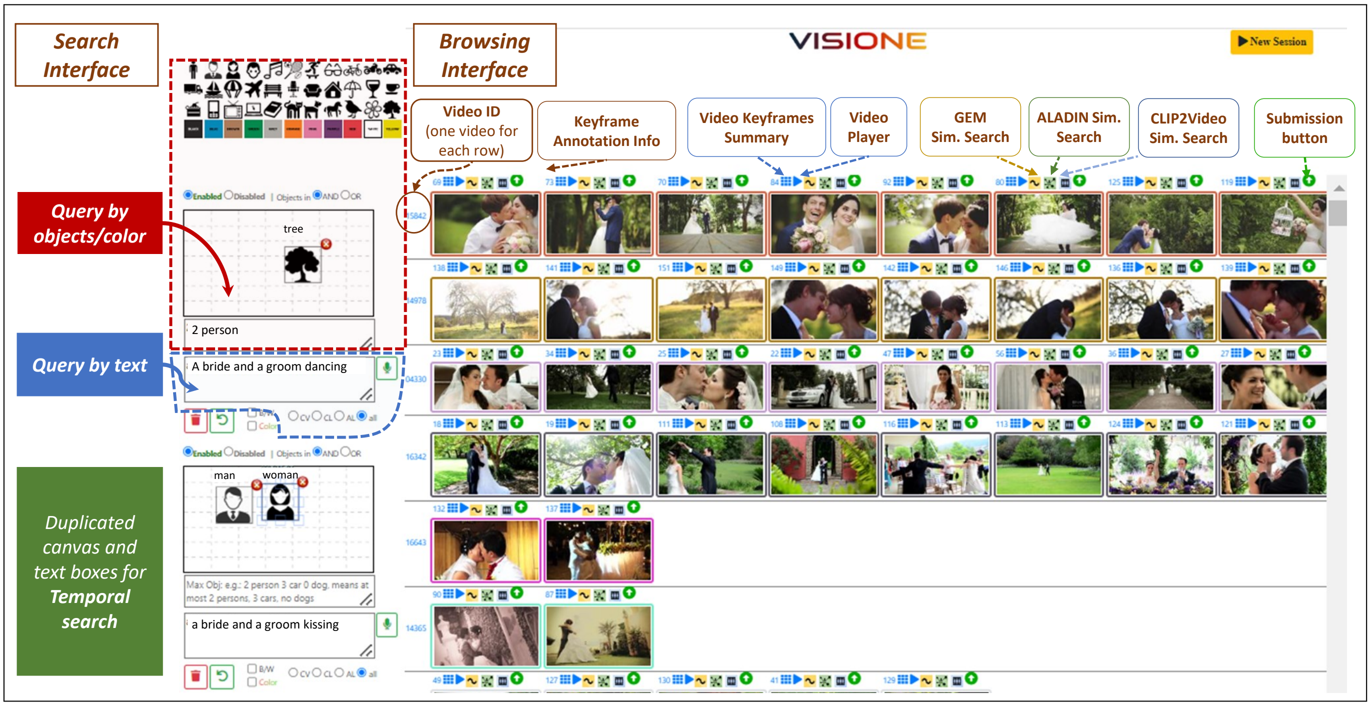
The VISIONE system allows users to perform temporal queries by specifying two different queries describing two temporally close scenes of a target video shot. A temporal quantization approach is used to search for videos that contain one keyframe satisfying the first query and another satisfying the second query. This involves dividing time into intervals of T seconds (e.g., $T = 3$) and processing the results of both queries independently to retain a single representative result (i.e., the one with the highest score) for each time interval and query. Result pairs that come from the same video and have a temporal distance smaller than a threshold (we used 12 seconds) are displayed to the user as results. Additionally, temporal quantization is used to present a limited number of result pairs from the same video, where only the result pair with the highest aggregated score in a specific time interval is considered.

3 THE NEW SIMPLIFIED USER INTERFACE

Figure 1 displays the user interface of the previous VISIONE release, which was participating at the VBS 2023. This interface was designed for expert users who were familiar with the system and its functionalities. It exposes all possible settings, including all supported cross-modal models and similarity search features. Moreover, it allows users to formulate text and object-based queries and combine both queries to search for a single scene or two temporally-close scenes. However, that interface may be overwhelming for new users due to the excessive number of options available. In most cases, a simple textual query or a temporal query with textual descriptions of two consecutive scenes is enough to find the desired shot. As a result, we have redesigned the user interface to make it more simple and easier to use.

Figure 2 displays the new user interface. The interface features a simple homepage that, drawing inspiration from major search engines like Google and Bing, displays only the system logo and a single search bar (top-left picture in Figure 2). Users can enter a textual description of the desired scene into the search bar (complex and detailed natural language queries are supported). Once a query is issued, the system promptly displays the relevant keyframes from the video database. Each row in the browsing interface corresponds to a video, and the first columns contain the most relevant results to the query according to our ranking model. Additionally, a second search bar is provided alongside the results to enable users to perform temporal queries. In other words, users can use this second

[1] https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K
[2] https://lucene.apache.org/
[3] https://github.com/facebookresearch/faiss



2 person

A bride and a groom dancing

Max Obj: e.g.: 2 person 3 car 0 dog, means at most 2 persons, 3 cars, no dogs

a bride and a groom kissing

Figure 1: Old VISIONE User Interface, used in the system release that participated at VBS2023.

input bar to specify what they want to happen a few seconds after the first searched scene (top-right picture in Figure 2).

Each search bar has two buttons (an X and a microphone). The X button clears the text. The microphone button allows users to dictate their textual query instead of typing it. If necessary, the spoken text is automatically translated into English and used as the query for the cross-modal search.

In the upper-left part of the interface, there are two more buttons: one is a restart button that enables resetting the whole interface and starting again from the home page; the other is the "*Advanced mode*" toggle button that activates the advanced interface. From the advanced interface (bottom-right picture in Figure 2), it is possible to obtain fine-grained control over the search capabilities. For example, it is possible to employ different models for running text-based queries (Clip, ClipLAION, ClipVideo, or all of them). By default, an aggregation of the three available methods is used. Furthermore, the advanced user interface allows users to perform also query by object- and color-based locations. Objects can be placed from a palette onto specific portions of the canvas, or a bounding box can be drawn and the class specified afterward (about 1,460 different objects are supported). Furthermore, there is a "*Max Obj.*" field where users can specify the maximum number of instances of a particular class that should be present in the target scene. The second canvas is for composing a temporal query. In summary, in the advanced mode, users can use the canvas-based querying system instead of relying solely on text input but still have the option to use the text input for finer control over the results and combine the various modalities with a temporal search. However, given the excellent performance of the cross-modal models employed in VISIONE, we expect that users will be able to fulfill their information needs without having to resort to the advanced mode of the user interface. Nonetheless, as users become more familiar with the system, they may desire to utilize or merge all of the search options provided by the system. Therefore, the dual mode of the interface has been developed to cater to this need.

Whether the advanced mode is on or off, the browsing component of the interface remains the same. Users can play a preview of the video with a mouse right-click or browse the entire video by clicking on the play icon. The matrix icon opens a new window containing all the indexed keyframes of the selected video. This visualization can be used by the users to quickly scan keyframes of the video in temporal order to see what happens before and after the selected one. During a competition, the user can submit a keyframe as a result directly by clicking on the green up arrow icon. Additionally, VISIONE allows users to find similar keyframes to a selected one by clicking on the tilde icon. In contrast to the previous release of VISIONE, the new version's similarity search button returns results obtained through a late fusion of three distinct similarity searches. Firstly, a visual similarity search based on GEM features [26] is employed. Secondly, a similarity search for video keyframes that are semantically similar to the input query based on ALADIN [21] features is conducted. Finally, a similarity search for video clips that are semantically similar to the input query based on CLIP2Video [10] features is performed. The new version, therefore, does not allow the user to select which model to use for the similarity search but uses all three models at the same time. This choice was intended to simplify the interface and its usability by novice users. Moreover, to aid users in remembering

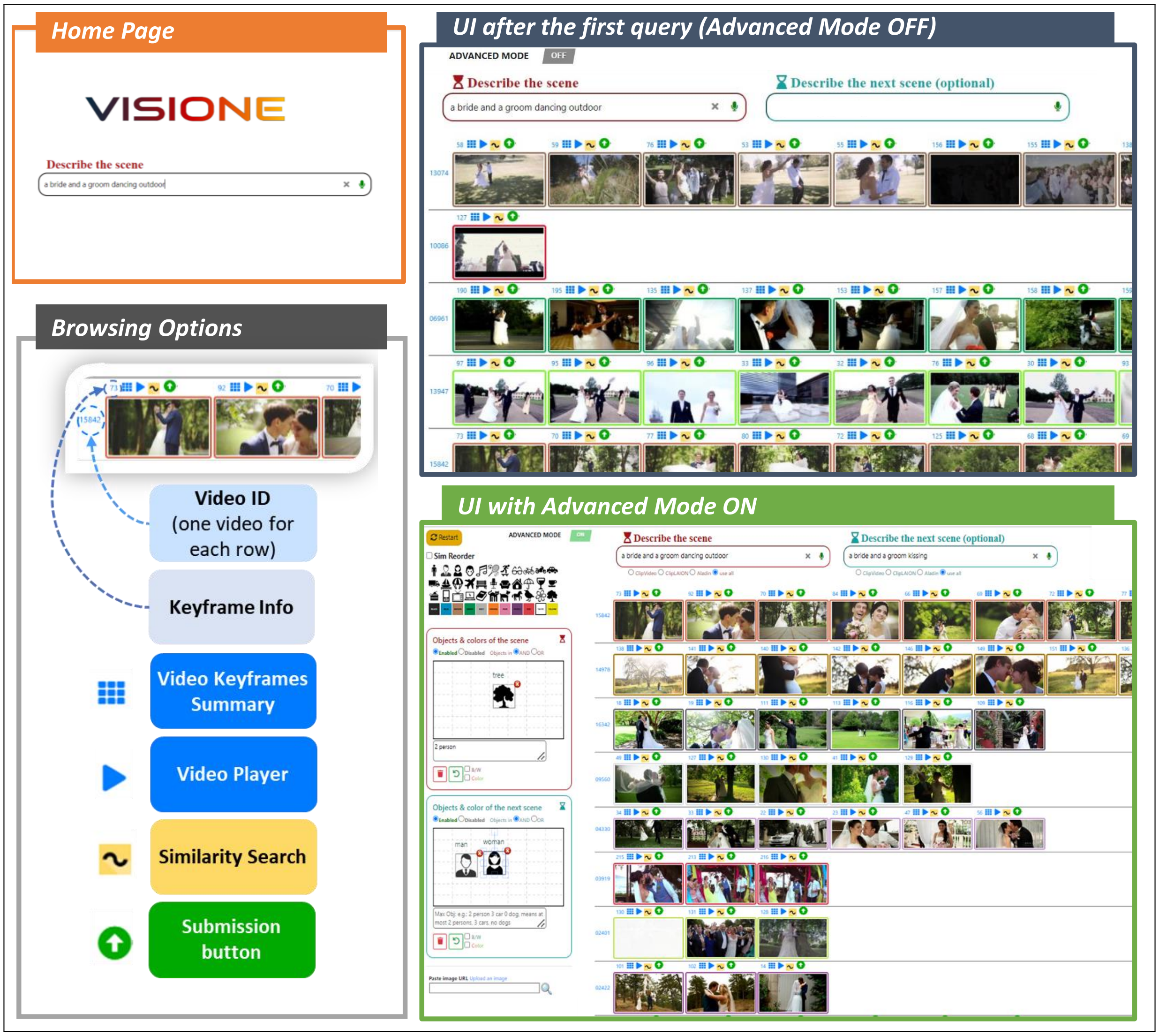

Home Page
VISIONE
Describe the scene
a bride and a groom dancing outdoor
Browsing Options
Video ID
(one video for each row)
Keyframe Info
Video Keyframes Summary
Video Player
Similarity Search
Submission button
UI after the first query (Advanced Mode OFF)
ADVANCED MODE
OFF
Describe the scene
a bride and a groom dancing outdoor
Describe the next scene (optional)
UI with Advanced Mode ON
Restart
ADVANCED MODE
ON
Sim Reorder
Describe the scene
a bride and a groom dancing outdoor
Describe the next scene (optional)
a bride and a groom kissing
Objects & colors of the scene
tree
2 person
Objects & color of the next scene
man
woman
Paste image URL Upload an image

Figure 2: New User Interface. The homepage (top-left figure) is quite simple and easy to use as it contains a single search bar where the user can enter a textual description of the desired scene. Once the query is entered, the system promptly displays search results along with a second input bar that can be used to perform a temporal search (top-right figure). A toggle button in the upper part of the UI can enable an "advanced search mode". The UI with the "advanced mode" turned ON is shown in the bottom-right figure. In that case, the user can access all the VISIONE functionalities, including the search based on object and color locations and the options on the cross-modal models to be used.

the functionality of each button, a text displaying the name of the function appears when the mouse cursor hovers over it.

4 CONCLUSIONS

This paper presented a revised version of the VISIONE video retrieval system that offers a wide range of search functionalities, including free text search, spatial color and object search, similarity search, and temporal search. The system leverages advanced indexing techniques and state-of-the-art AI technology to ensure scalability and effectiveness. The redesigned search and browsing interface aims to improve the system's usability.

ACKNOWLEDGMENTS

This work was partially funded by AI4Media - A European Excellence Centre for Media, Society, and Democracy (EC, H2020 n. 951911) and National Centre for HPC, Big Data and Quantum Computing – HPC (CUP B93C22000620006).

REFERENCES

[1] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. 2019. VISIONE at VBS2019. In *MultiMedia Modeling*. Springer, 591–596.

[2] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. 2021. The VISIONE video search system: exploiting off-the-shelf text search engines for large-scale video retrieval. *Journal of Imaging* 7, 5 (2021), 76.

[3] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2022. VISIONE at Video Browser Showdown 2022. In *MultiMedia Modeling*. Springer International Publishing, Cham, 543–548.

[4] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2023. VISIONE: A Large-Scale Video Retrieval System with Advanced Search Functionalities *(ICMR '23)*. Association for Computing Machinery, New York, NY, USA, 649–653.

[5] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2023. VISIONE at Video Browser Showdown 2023. In *MultiMedia Modeling*. Springer International Publishing, Cham, 615–621.

[6] Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2021. VISIONE at Video Browser Showdown 2021. In *International Conference on Multimedia Modeling*. Springer, 473–478.

[7] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. 2019. Large-scale instance-level image retrieval. *Information Processing & Management* (2019), 102100.

[8] Robert Benavente, Maria Vanrell, and Ramon Baldrich. 2008. Parametric fuzzy sets for automatic color naming. *JOSA A* 25, 10 (2008), 2582–2593.

[9] Fabio Carrara, Lucia Vadicamo, Claudio Gennaro, and Giuseppe Amato. 2022. Approximate Nearest Neighbor Search on Standard Search Engines. In *Similarity Search and Applications*, Tomáš Skopal, Fabrizio Falchi, Jakub Lokoč, Maria Luisa Sapino, Ilaria Bartolini, and Marco Patella (Eds.). Springer International Publishing, Cham, 214–221.

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

[11] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[12] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *International Conference on Multimedia Retrieval (ICMR'22). ACM*.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[14] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, Luca Rossetto, et al. 2022. 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* 11, 1 (2022), 1–18.

[15] Nhat Hoang-Xuan, E-Ro Nguyen, Thang-Long Nguyen-Ho, Minh-Khoi Pham, Quang-Thuc Nguyen, Hoang-Phuc Trang-Trung, Van-Tu Ninh, Tu-Khiem Le, Cathal Gurrin, and Minh-Triet Tran. 2023. V-FIRST 2.0: Video Event Retrieval with Flexible Textual-Visual Intermediary for VBS 2023. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Springer, 652–657.

[16] Jakub Lokoč, Stelios Andreadis, Werner Bailer, Aaron Duane, Cathal Gurrin, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peška, Luca Rossetto, Loris Sauter, Konstantin Schall, Klaus Schoeffmann, Omar Shahbaz Khan, Florian Spiess, Lucia Vadicamo, and Stefanos Vrochidis. 2023. Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th VBS. *Multimedia Systems* (24 Aug 2023). https://doi.org/10.1007/s00530-023-01143-5

[17] Jakub Lokoč, Werner Bailer, Kai Uwe Barthel, Cathal Gurrin, Silvan Heller, Björn Þór Jónsson, Ladislav Peška, Luca Rossetto, Klaus Schoeffmann, Lucia Vadicamo, et al. 2022. A task category space for user-centric comparative multimedia search evaluations. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*. Springer, 193–204.

[18] Jakub Lokoč, Werner Bailer, Klaus Schoeffmann, Bernd Muenzer, and George Awad. 2018. On influential trends in interactive video retrieval: Video Browser Showdown 2015-2017. *IEEE Transactions on Multimedia* 20, 12 (Dec 2018), 3361–3376.

[19] Jakub Lokoč, Patrik Veselỳ, František Mejzlík, Gregor Kovalčík, Tomáš Souček, Luca Rossetto, Klaus Schoeffmann, Werner Bailer, Cathal Gurrin, Loris Sauter, et al. 2021. Is the Reign of Interactive Search Eternal? Findings from the Video Browser Showdown 2020. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 3 (2021), 1–26.

[20] Jakub Lokoč, Zuzana Vopálková, Patrik Dokoupil, and Ladislav Peška. 2023. Video Search with CLIP and Interactive Text Query Reformulation. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Springer, 628–633.

[21] Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. 2022. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. *arXiv preprint arXiv:2207.14757* (2022).

[22] Thao-Nhu Nguyen, Bunyarit Puangthamawathanakun, Annalina Caputo, Graham Healy, Binh T Nguyen, Chonlameth Arpnikanondt, and Cathal Gurrin. 2023. VideoCLIP: An Interactive CLIP-based Video Retrieval System at VBS2023. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Springer, 671–677.

[23] Nick Pantelidis, Stelios Andreadis, Maria Pegia, Anastasia Moumtzidou, Damianos Galanopoulos, Konstantinos Apostolidis, Despoina Touska, Konstantinos Gkountakos, Ilias Gialampoukidis, Stefanos Vrochidis, et al. 2023. VERGE in VBS 2023. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Springer, 658–664.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. *Robust speech recognition via large-scale weak supervision*. Technical Report. Tech. Rep., Technical report, OpenAI.

[26] J. Revaud, J. Almazan, R.S. Rezende, and C.R. de Souza. 2019. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In *International Conference on Computer Vision*. IEEE, 5106–5115.

[27] L. Rossetto, R. Gasser, J. Lokoc, W. Bailer, K. Schoeffmann, B. Muenzer, T. Soucek, P. A. Nguyen, P. Bolettieri, A. Leibetseder, and S. Vrochidis. 2020. Interactive Video Retrieval in the Age of Deep Learning - Detailed Evaluation of VBS 2019. *IEEE Transactions on Multimedia* (2020), 1–1.

[28] Loris Sauter, Ralph Gasser, Silvan Heller, Luca Rossetto, Colin Saladin, Florian Spiess, and Heiko Schuldt. 2023. 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*. Springer, 646–651.

[29] Konstantin Schall, Nico Hezel, Klaus Jung, and Kai Uwe Barthel. 2023. Vibro: Video Browsing with Semantic and Visual Image Embeddings. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Springer, 665–670.

[30] Klaus Schoeffmann, Daniela Stefanics, and Andreas Leibetseder. 2023. diveXplore at the Video Browser Showdown 2023. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Springer, 684–689.

[31] Florian Spiess, Silvan Heller, Luca Rossetto, Loris Sauter, Philipp Weber, and Heiko Schuldt. 2023. Traceable Asynchronous Workflows in Video Retrieval with vitrivr-VR. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part I*. Springer, 622–627.

[32] Ly-Duyen Tran, Manh-Duy Nguyen, Duc-Tien Dang-Nguyen, Silvan Heller, Florian Spiess, Jakub Lokoč, Ladislav Peška, Thao-Nhu Nguyen, Omar Shahbaz Khan, Aaron Duane, et al. 2023. Comparing Interactive Retrieval Approaches at the Lifelog Search Challenge 2021. *IEEE Access* (2023).

[33] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. 2009. Learning color names for real-world applications. *IEEE Transactions on Image Processing* 18, 7 (2009), 1512–1523.

[34] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. 2021. VarifocalNet: An IoU-aware Dense Object Detector. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.