VISIONE 5.0: Enhanced User Interface and AI Models for VBS2024

Giuseppe Amato ⓘ, Paolo Bolettieri ⓘ, Fabio Carrara ⓘ, Fabrizio Falchi ⓘ,
Claudio Gennaro ⓘ, Nicola Messina ⓘ, Lucia Vadicamo ⓘ, and Claudio Vairo ⓘ

CNR-ISTI, Via G. Moruzzi 1, 56124 Pisa, Italy
`name.surname@isti.cnr.it`

Abstract. In this paper, we introduce the fifth release of VISIONE, an advanced video retrieval system offering diverse search functionalities. The user can search for a target video using textual prompts, drawing objects and colors appearing in the target scenes in a canvas, or images as query examples to search for video keyframes with similar content. Compared to the previous version of our system, which was runner-up at VBS 2023, the forthcoming release, set to participate in VBS 2024, showcases a refined user interface that enhances its usability and updated AI models for more effective video content analysis.

Keywords: Information Search and Retrieval · Content-based video retrieval · Video search · Surrogate Text Representation · Multi-modal Retrieval · Cross-modal retrieval

1 Introduction

Video Browser Showdown (VBS) [15,13,17] is an important international competition for interactive video search [16], which is held annually at the International Conference on MultiMedia Modeling (MMM) since 2012. To date, it has comprised three different search tasks, namely *visual Known-Item Search* (KIS-V), *textual Known-Item Search* (KIS-T) and *Ad-hoc Video Search* (AVS). In KIS-V and KIS-T, the user must retrieve a specific segment of a video starting from, respectively, a visual or textual hint. The visual cue is given by playing the target video segment on a large screen visible to all competition participants (or on a browser using the DRES evaluation server [23], to which all teams' systems are connected). The textual hint is the description of the video segment, which is gradually extended during the task time with more details useful to identify the target video segment uniquely. In the AVS, the user has to retrieve as many as possible video segments matching a textual description provided.

The size or diversity of datasets used in the competition has increased over the years to make the competition more challenging. Last year, two datasets were used: a large dataset (V3C1+V3C2) [24], composed of 17,235 diverse videos for a total duration of 2,300 hours, and the Marine Video Kit (MVK) dataset [29], which is composed of 1,372 highly redundant videos taken from moving cameras in underwater environments (scuba diving).

Usually, the competition is organized in expert and novice sessions. In the novice session, volunteers from the audience will be recruited to solve tasks with the search system after a brief tutorial on how to use it. Therefore, one of the focuses when implementing such systems is also to make them easy for a non-expert user to use.

There will be two main novelties in the 2024 edition of the competition. First, another dataset has been added as a target dataset for queries, the VBSLHE dataset. It is composed of 75 videos capturing laparoscopic gynecology surgeries. This dataset is particularly challenging because videos are very similar to each other, and it is difficult for non-medic users to understand what is happening in the scene and to submit more precise queries. The other novelty is the introduction of the Question Answering (Q/A) task, where the user is asked to submit textual answers to questions regarding the collection (e.g., "What airline name appears most frequently in the dataset?") instead of submitting video segments.

Our video retrieval system VISIONE [3,7,2,1] has participated in four editions of the VBS competition since 2019, skipping the 2020 edition. In the 2023 competition, with 13 systems participating, our system ranked first in the KIS visual task and second in the entire competition, behind Vibro [25] and ahead of other very effective systems, including Vireo [19], vitrivrVR [28], and CVHunter [18]. This paper presents the fifth version of VISIONE, which includes some changes compared to the previous versions. These changes are mainly focused on a new simplified and easy-to-use interface described in Section 3. In addition, it is worth noting that recently, we have made available the features we extracted from all the AI models we exploit in our system to the scientific community through several Zenodo repositories [6,5,4].

2   System Overview

VISIONE provides a set of search functionalities that empower the user to search specific video segments through text and visual queries, which can also be combined into a temporal search. The system provides *free text search*, *spatial color and object search*, and *visual-semantic similarity search*. For a comprehensive understanding of the system's architecture and an overview of the user interface, please refer to our previous publications, which describe these details [3,1].

In pursuit of robust and efficient free text and semantic similarity search, we leverage three cross-modal feature extractors, each driven by pre-trained models. OpenCLIP ViT-L/14 [14,22] model pre-trained on LAION-2B dataset [27] (in the following we refer to this model as ClipLAION), CLIP2Video [12], and ALADIN [20]. By default, a combination of all three models is used when performing a free text search, which is achieved by merging the results with a late fusion approach. However, if needed, it is possible to select one of the aforementioned models by selecting the desired model in the dedicated radio button in the UI. Three models are also used for the object detection: VfNet trained on the COCO dataset, Mask R-CNN trained on the LVIS dataset, and a Faster R-CNN+Inception ResNet V2 trained on Open Images V4 (see [3]).
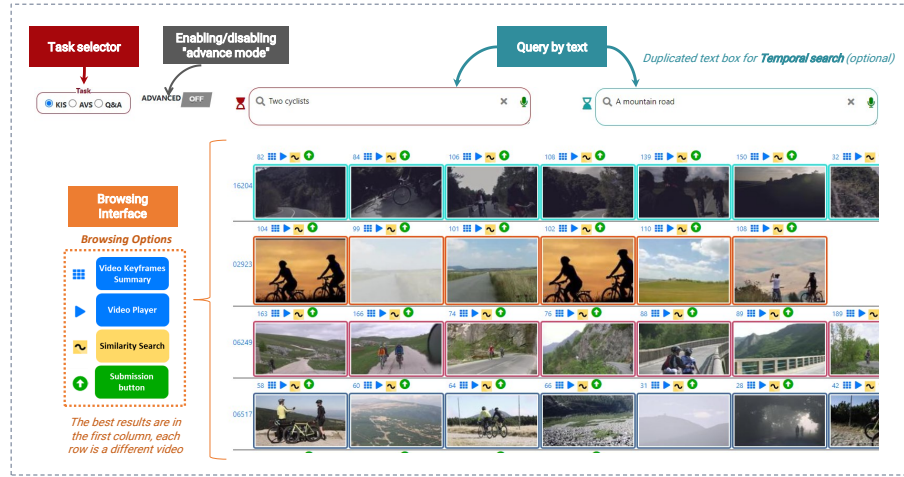
Fig. 1: New simplified user interface with textual, temporal and similarity search. When clicking on "Advanced Mode", the user can access advanced search options, including object search and the capability to choose a specific AI model for cross-modal searches.

The visual similarity search leverages the comparative analysis of the extracted DINOv2 [21] features that have been proven to be effective for several image-level visual tasks (image classification, instance retrieval, video understanding). Finally, the spatial color-based search is enabled by the annotations extracted using the two chip-based color naming techniques (see [3]).

For our indexing strategy, we leverage two different techniques: the Facebook FAISS library[1] is used to store and access both CLIP2Video and ClipLAION features. The second index is tailored to store all other descriptors and can be effortlessly queried using Apache Lucene[2]. Notably, for indexing all the extracted descriptors with Lucene, we devised specialized text encodings based on the Surrogate Text Representations (STRs) approach described in [1,8,10,9].

3   Recent Changes to the VISIONE System

In this section, we present the changes that we introduced in VISIONE with respect to the system version participating in the last VBS competition.

Simplified Interface. We added a simplified version of the user interface to help novice users to perform searches. The new simplified interface, now the default

---

[1] https://github.com/facebookresearch/faiss
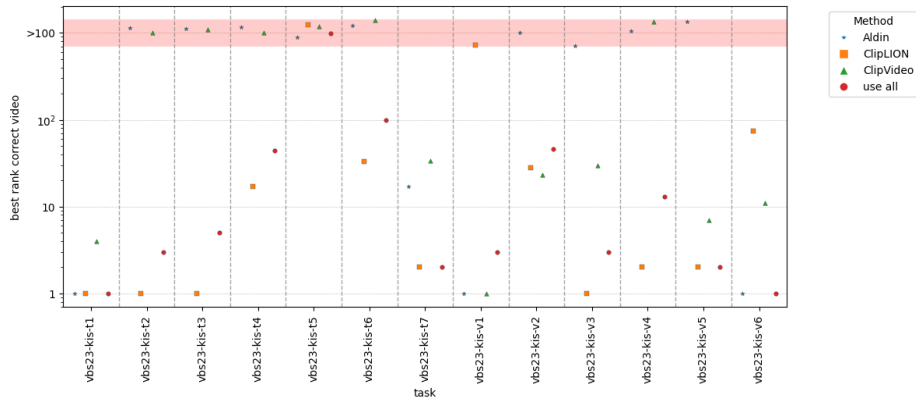[2] https://lucene.apache.org/

Fig. 2: Distributions of the best ranks of correct video per task and method. We used all the textual queries issued by the two users of the VISIONE system during the last VBS2023 competition. Note that only text queries made on the V3C dataset were considered, excluding the combination with other search features, e.g., temporal search.

one, shows basic search functionalities (see Figure 1). In the text boxes, it is possible to submit a query by text that exploits a combination of the three networks we used to extract features from the videos, namely ALADIN, Clip2Video, and ClipLAION. The second text box (the green one on the right in Figure 1) is optional and can be used to perform a temporal query, in which the user can describe a scene occurring after the scene described in the primary red text box on the left. On top of each returned result, there is the ID of the shot that the frame refers to and four buttons that allow some browsing options. From left to right, the grid button allows to show all the keyframes of that video, the play button starts the playback of the video, the tilde button performs an aggregated visual similarity search to find frames visually similar to the selected one, and the white arrow in the green circle button submits the corresponding frame to DRES.

Regarding the similarity search, in the previous versions of VISIONE, there was the possibility to perform three different visual similarity searches, one for each network used to extract visual features, namely DINOv2, ALADIN, and Clip2Video. In this version, we use an aggregated method of similarity search that relies on late fusion to combine the ranked results obtained from the single approaches. The late fusion employed is the Reciprocal Rank Fusion [11], with an additional enhancement on the top-5 results from each ranked list. This is the same heuristic used to combine results in the free text search. Based on our preliminary experiments using the VISIONE logs collected during VBS2023 (see Figure 2), while ClipLION exhibits the overall best performance, there are instances in which other methods correctly identify the video with a better ranking (e.g., tasks vbs23-kis-v1 and vbs23-kis-v6). Our late fusion strategy doesn't con-

sistently achieve the best rankings, however, it enables us to strike a favorable compromise among the various methods employed, allowing the user not to be forced to choose one of the search methods.

An "Advanced mode" button in the top-left corner of the interface allows enabling an advanced version of the interface with all the search functionalities available. In this modality, the user can select a specific model to be used for text search. In the new interface, we also added the possibility to click on a selected frame to get a zoomed view of that particular frame.

We presented this new interface to the "Interactive Video Retrieval for Beginners" special session at the 20th International Conference on Content-Based (CBMI 2023) [7]. This occasion has been very useful in collecting valuable feedback from the novice users who actually utilized our system. We will consider this feedback as we work to enhance the usability of VISIONE further.

Finally, we observe that in the top-left corner of the interface, it is possible to switch between the different tasks of the competition (KIS, AVS, and Q&A), and between the different datasets (V3C, MVK, VBSLHE). This triggers some ad-hoc modifications to the user interface to address functionalities tailored for that particular task or dataset, for example, a different object palette for different datasets, the addition of a text box for the Q&A task, and a different browsing behavior for the AVS task. To search the VBSLHE dataset, we used BiomedCLIP [30], a biomedical vision-language foundation model that is pretrained on PMC-15M [31], a dataset of 15 million figure-caption pairs extracted from biomedical research articles in PubMed Central, using contrastive learning.

Keyboard shortcuts. AVS is one of the tasks where VISIONE performed worst in the previous competitions. Inspired by lifeXplore system [26], we tried to improve the usability of the system in this task by adding some keyboard shortcuts. In particular, it is possible to browse the frames resulting from a search using the arrow keys and to submit the current selected frame by pressing the "s" key. This makes selecting and submitting frames for the AVS task quicker than using the mouse.

Features of all datasets available in Zenodo. We extracted and made publicly available on Zenodo [6,5,4] the features and annotations extracted from all the dataset used in the completion (V3C, MVK, VBSLHE) for all the three multimodal models used in our system (ALADIN, Clip2Video and ClipLAION) and for the object detectors (VfNet, Mask R-CNN, and Faster R-CNN+Inception ResNet V2). We also provide a Python script to extract the keyframes used by the VISIONE system for which features and annotations are provided.

Question & Answering Task. To date, VISIONE does not allow automatic processing for the new question and answering task. We simply added a dedicated textbox to the UI by which the user can manually submit the answer when found. To do that, the user should rephrase the questions into text queries suitable for finding a set of relevant results, which she/he should then review to locate the desired answer.

4   Conclusions and Future Works

The paper introduces the fifth version of the VISIONE, a video retrieval system that offers advanced functionalities to search for specific video segments using free text search, spatial color and object search, and visual-semantic similarity search. Compared to the previous version, the new release for VBS 2024 features a simplified user interface, which could be helpful for novice users, and enhanced usability through keyboard shortcuts. It also includes the public release of features and annotations extracted from the datasets used for the competition. The 2024 VBS competition introduces new challenges, such as integrating the VBSLHE dataset and a Question Answering (Q/A) task. VISIONE addresses these challenges by improving its capabilities, incorporating new AI models, and enhancing its natural language processing capabilities.

We plan to release the code of our system as open source soon. This release will encompass not only the core search engine but also the user interface (UI). By making our system open source, we aim to encourage broader usage and invite collaboration for further system development. An example of a potential direction we envision is creating a Virtual Reality (VR) user interface for the system.

In our future work, we aim to automate the execution of the Q&A task by implementing natural language processing and question-answering techniques, enabling the user to directly input her/his questions and obtain a pool of potential textual answers.

References

1. Amato, G., Bolettieri, P., Carrara, F., Debole, F., Falchi, F., Gennaro, C., Vadicamo, L., Vairo, C.: The VISIONE video search system: exploiting off-the-shelf text search engines for large-scale video retrieval. Journal of Imaging **7**(5), 76 (2021)
2. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: Visione: A large-scale video retrieval system with advanced search functionalities. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. pp. 649–653 (2023)

3. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE at Video Browser Showdown 2023. In: MultiMedia Modeling. pp. 615–621. Springer (2023)
4. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE Feature Repository for VBS: Multi-Modal Features and Detected Objects from LapGyn100 Dataset (Oct 2023). https://doi.org/10.5281/zenodo.10013328
5. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE Feature Repository for VBS: Multi-Modal Features and Detected Objects from MVK Dataset (Sep 2023). https://doi.org/10.5281/zenodo.8355037
6. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE Feature Repository for VBS: Multi-Modal Features and Detected Objects from V3C1+V3C2 Dataset (Jul 2023). https://doi.org/10.5281/zenodo.8188570
7. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE for newbies: an easier-to-use video retrieval system. In: Proceedings of the 20th International Conference on Content-based Multimedia Indexing. Association for Computing Machinery (2023)
8. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Vadicamo, L.: Large-scale instance-level image retrieval. Information Processing & Management **57**(6), 102100 (2020)
9. Carrara, F., Gennaro, C., Vadicamo, L., Amato, G.: Vec2Doc: Transforming Dense Vectors into Sparse Representations for Efficient Information Retrieval. In: Similarity Search and Applications. Springer, Cham (2023)
10. Carrara, F., Vadicamo, L., Gennaro, C., Amato, G.: Approximate Nearest Neighbor Search on Standard Search Engines. In: Similarity Search and Applications. pp. 214–221. Springer, Cham (2022)
11. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 758–759 (2009)
12. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
13. Heller, S., Gsteiger, V., Bailer, W., Gurrin, C., Jónsson, B.Þ., Lokoč, J., 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 **11**(1), 1–18 (2022)
14. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., et al.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773
15. Lokoč, J., Andreadis, S., Bailer, W., Duane, A., Gurrin, C., Ma, Z., Messina, N., Nguyen, T.N., Peška, L., Rossetto, L., et al.: Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th vbs. Multimedia Systems pp. 1–24 (2023)
16. Lokoč, J., Bailer, W., Barthel, K.U., Gurrin, C., Heller, S., Jónsson, B.Þ., Peška, L., Rossetto, L., Schoeffmann, K., Vadicamo, L., et al.: A task category space for user-centric comparative multimedia search evaluations. In: MultiMedia Modeling. pp. 193–204. Springer (2022)
17. Lokoč, J., Veselỳ, P., Mejzlík, F., Kovalčík, G., Souček, T., Rossetto, L., Schoeffmann, K., Bailer, W., Gurrin, C., Sauter, L., et al.: Is the reign of interactive search

eternal? findings from the video browser showdown 2020. ACM Transactions on Multimedia Computing, Communications, and Applications **17**(3), 1–26 (2021)

18. Lokoč, J., Vopálková, Z., Dokoupil, P., Peška, L.: Video search with clip and interactive text query reformulation. In: MultiMedia Modeling. pp. 628–633. Springer (2023)

19. Ma, Z., Wu, J., Loo, W., Ngo, C.W.: Reinforcement learning enhanced pichunter for interactive search. In: MultiMedia Modeling (2023)

20. Messina, N., Stefanini, M., Cornia, M., Baraldi, L., Falchi, F., Amato, G., Cucchiara, R.: Aladin: distilling fine-grained alignment scores for efficient image-text matching and retrieval. In: Proceedings of the 19th International Conference on Content-based Multimedia Indexing. pp. 64–70 (2022)

21. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th Intl. Conference on Machine Learning, ICML 2021. pp. 8748–8763. PMLR (2021)

23. Rossetto, L., Gasser, R., Sauter, L., Bernstein, A., Schuldt, H.: A system for interactive multimedia retrieval evaluations. In: MultiMedia Modeling. pp. 385–390. Springer (2021)

24. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3c–a research video collection. In: MultiMedia Modeling. pp. 349–360. Springer (2019)

25. Schall, K., Hezel, N., Jung, K., Barthel, K.U.: Vibro: Video browsing with semantic and visual image embeddings. In: MultiMedia Modeling. pp. 665–670. Springer (2023)

26. Schoeffmann, K.: lifexplore at the lifelog search challenge 2023. In: Proceedings of the 6th Annual ACM Lifelog Search Challenge. pp. 53–58 (2023)

27. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)

28. Spiess, F., Heller, S., Rossetto, L., Sauter, L., Weber, P., Schuldt, H.: Traceable Asynchronous Workflows in Video Retrieval with vitrivr-VR. In: MultiMedia Modeling. vol. 13833, pp. 622–627. Springer (2023)

29. Truong, Q.T., Vu, T.A., Ha, T.S., Lokoč, J., Tim, Y.H.W., Joneja, A., Yeung, S.K.: Marine video kit: A new marine video dataset for content-based analysis and retrieval. In: MultiMedia Modeling. Springer (2023)

30. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M., Naumann, T., Poon, H.: Large-scale domain-specific pretraining for biomedical vision-language processing (2023). https://doi.org/10.48550/ARXIV.2303.00915

31. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M.P., Naumann, T., Poon, H.: Large-scale domain-specific pretraining for biomedical vision-language processing (2023)