VISIONE at Video Browser Showdown 2023

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

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

Abstract. In this paper, we present the fourth release of VISIONE, a tool for fast and effective video search on a large-scale dataset. It includes several search functionalities like text search, object and color-based search, semantic and visual similarity search, and temporal search. VISIONE uses ad-hoc textual encoding for indexing and searching video content, and it exploits a full-text search engine as search backend. In this new version of the system, we introduced some changes both to the current search techniques and to the user interface.

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

1 Introduction

Video Browser Showdown (VBS) [19,14,11] is an annually-held international competition for video search on a large-scale dataset (V3C1 + V3C2) composed of 17,235 videos for a total duration of 2300 hours [20]. It consists of three different tasks: visual and textual known-item search (KIS) and ad-hoc video search (AVS). In the 2022 competition, our system VISIONE [3] ranked first in the KIS visual task, and third in the entire competition, behind Vibro [12] and CVHunter [13] ranked, respectively, first and second. In the 2023 edition of the competition, some important changes will be introduced. The duration of the target scene to be found in the KIS tasks will be reduced from 20 seconds to only 3 seconds. Moreover, there will be a dedicated session issuing tasks in the Marine Video Kit dataset [21], which is composed of highly redundant videos taken from moving cameras in underwater environments.

In this paper we present the fourth version of VISIONE [1,7,2,4,3] which includes important changes compared to the previous version. We implemented a new model for our text-to-image retrieval tool, called ALADIN [16], which replaces the TERN feature [15] used last year. In this version, we also included a text-to-video retrieval tool exploiting the state-of-the-art CLIP2Video [9] network. We improved our web interface to speed up the performance in AVS tasks by adding a mechanism for quickly selecting the most similar frames for each video. Finally, we implemented a clustering method to reduce the number of frames shown in the result set so that it is easier for the user to find the target video. All these changes will be described in more detail in Section 3.

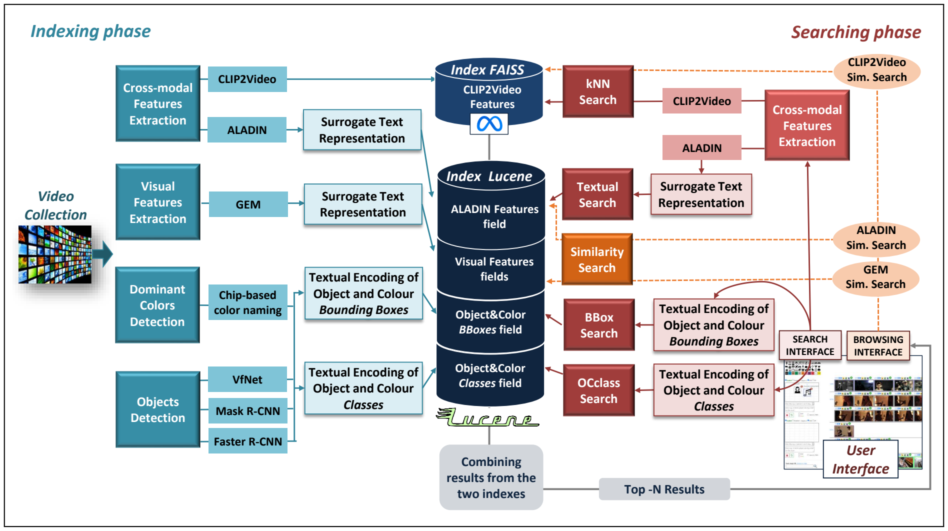

Indexing phase
Searching phase
Video Collection
Cross-modal Features Extraction
CLIP2Video
ALADIN
Surrogate Text Representation
Visual Features Extraction
GEM
Surrogate Text Representation
Dominant Colors Detection
Chip-based color naming
Textual Encoding of Object and Colour Bounding Boxes
Objects Detection
VfNet
Mask R-CNN
Faster R-CNN
Textual Encoding of Object and Colour Classes
Index FAISS
CLIP2Video Features
Index Lucene
ALADIN Features field
Visual Features fields
Object&Color BBoxes field
Object&Color Classes field
Lucene
Combining results from the two indexes
Top -N Results
kNN Search
CLIP2Video
Cross-modal Features Extraction
ALADIN
Textual Search
Surrogate Text Representation
Similarity Search
BBox Search
Textual Encoding of Object and Colour Bounding Boxes
OCclass Search
Textual Encoding of Object and Colour Classes
CLIP2Video Sim. Search
ALADIN Sim. Search
GEM Sim. Search
SEARCH INTERFACE
BROWSING INTERFACE
User Interface

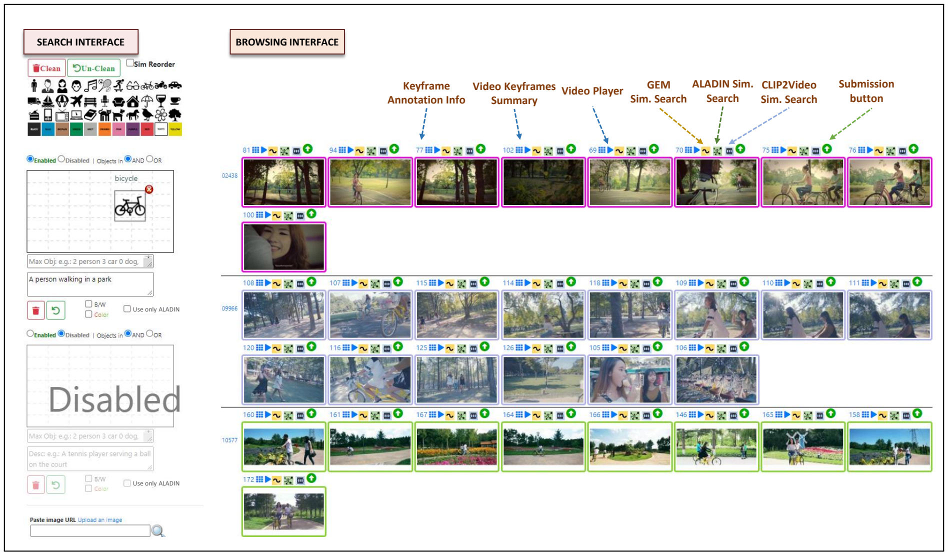
Fig. 1: VISIONE System Architecture

2 System Overview

VISIONE integrates several search functionalities that allow a user to search for a target video segment by formulating textual and visual queries, which can be also combined with a temporal search. In particular it supports *free text search*, *spatial color and object search*, *visual similarity search*, and *semantic similarity search*. The system architecture is summarized in Figure 1, while a screenshot of the user interface is shown in Figure 2.

To support the free text search and the semantic similarity search, we employ two cross-modal feature extractors based on, respectively, CLIP2Video [9] and ALADIN [16] pre-trained models. For the object detection, we use three models: VfNet[23] (trained on COCO dataset), Mask R-CNN [10] (trained on LVIS dataset), and a Faster R-CNN+Inception ResNet V2[1] (trained on the Open Images V4). The color annotation process relies on two chip-based color naming techniques [22,6]. Finally, the visual similarity search is based on comparing GEM [18] features. We employ two indexes: the first to store the CLIP2Video features (searched using the Facebook FAISS library[2]), and the second to store all the other descriptors (searched using Apache Lucene[3]). Note that to index the extracted descriptors with Lucene, we designed special text encodings, based on the Surrogate Text Representations (STRs) approach [2,5,8].

[1] http://tfhub.dev/google/faster_rcnn/openimages_v4/inception_resnet_v2/1
[2] https://github.com/facebookresearch/faiss
[3] https://lucene.apache.org/



A person walking in a park

Fig. 2: User Interface

3 Recent Changes to the VISIONE System

Compared to last year's system description [3], we modified some features used for object, similarity, and text search, as described below[4].

Using ALADIN for text-to-image retrieval. We developed a new cross-modal retrieval deep neural network, called ALADIN (ALign And DIstill Network) [16]. ALADIN first produces high-effective scores by aligning at fine-grained level images and texts. Then, it learns a shared embedding space – where an efficient kNN search can be performed – by distilling the relevance scores obtained from the fine-grained alignments. We empirically found that this network is able to compete with state-of-the-art vision-language Transformers while being almost 90 times faster at inference time.

ALADIN visual features can also be used to perform an image-to-image similarity search, which showed remarkable performance in semantic image retrieval.

Using CLIP2Video for text-to-video retrieval. In order to deeply understand videos, in particular temporal correlations and actions among multiple frames of a shot, we use CLIP2Video [9], which is one of the state-of-the-art networks for text-to-video retrieval. We re-engineered the code for easily extracting fixed-sized descriptors for texts and images that can be compared with cosine similarity. However, we found some problems in post-processing these features using

[4] Please note that some of these changes were already integrated in VISIONE some weeks before the last VBS competition

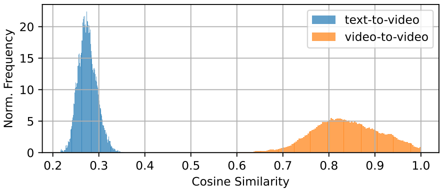


Fig. 3: Distribution of cosine similarities between text-video and video-video CLIP2Video features.

our STR representation for textual-based indexing [5]. In particular, looking at Figure 3, we noticed that the distribution of the cosine similarities of the CLIP2Video features has a very low mean value in the text-to-image cross-modal setup. This may happen if element-wise products underlying the dot-product computation have a negative sign, which in turn implies that there could be a lot of mixed-sign factors. This is a bad scenario for the STR representation, given that the CReLU operation at the core of the STR method zeroes out the contribution from mixed-sign factors. Therefore, for the CLIP2Video features, the approximated cosine similarity computed in the STR representation badly approximates the original one. For these reasons, we indexed and searched these cross-modal features with FAISS, using an exact search and an 8-bit scalar quantization for reducing the index size in memory.

The visual features extracted using CLIP2Video are also employed for a semantic reverse video search, where a video segment displayed in the results can be used as a query to search other video clips semantically similar to it.

Improvements in the Browsing Interface. To improve the user's browsing experience, we included the possibility of displaying a short preview of a video clip by right-clicking on one of the results displayed in the user interface. We also introduced multiple selections of frames to submit during AVS tasks and the ability to submit a given instant of a video directly from the videoplayer.

Object search. We used three object detectors trained on three different datasets (COCO, LVIS, and Open Images V4), which have different classes. We built a mapping of these classes using a semi-automatic procedure in order to have a unique final list of 1,460 classes [5]. We also generated a hierarchy for each class, using wordnet[6], which is used for query expansion both at index time and at runtime.

Planned changes. The current display of results is based on a grouping of images by video, and for each video, only the 20 keyframes with higher scores are displayed to the user. One of the main drawbacks of such visualization is

[5] `https://doi.org/10.5281/zenodo.7194300`

[6] https://wordnet.princeton.edu/

the presence of many near duplicate keyframes, which burdens the visualization without adding distinctive information. We, therefore, plan to use hierarchical clustering techniques to improve the result visualization.

Moreover, we plan to exploit the Whisper model [17] to add a speech-to-text functionality that would facilitate issuing a textual query by dictating it to the system instead of typing it.

4 Conclusions and Future Work

This paper presents the novelties introduced in the VISIONE system for participating to the next edition of the Video Browser Showdown. To further improve the system, in the future we would like to exploit relevant feedback techniques, which may speed up and increase the number of correct results during AVS tasks. Moreover, we plan to investigate novel STR encoding techniques to effectively transform CLIP2Video features into textual documents. This would allow us to get rid of the FAISS index and be able to use a single scalable full-text search engine for indexing all the different descriptors.

Acknowledgements

This work was partially funded by AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911) and INAROS, CNR4C program (Tuscany POR FSE CUP B53D21008060008).

References

1. Amato, G., Bolettieri, P., Carrara, F., Debole, F., Falchi, F., Gennaro, C., Vadicamo, L., Vairo, C.: VISIONE at VBS2019. In: MultiMedia Modeling. pp. 591–596. Springer (2019)
2. 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)
3. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE at Video Browser Showdown 2022. In: MultiMedia Modeling. pp. 543–548. Springer International Publishing, Cham (2022)
4. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE at Video Browser Showdown 2021. In: International Conference on Multimedia Modeling. pp. 473–478. Springer (2021)
5. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Vadicamo, L.: Large-scale instance-level image retrieval. Information Processing & Management p. 102100 (2019)
6. Benavente, R., Vanrell, M., Baldrich, R.: Parametric fuzzy sets for automatic color naming. JOSA A **25**(10), 2582–2593 (2008)
7. Bolettieri, P., Carrara, F., Debole, F., Falchi, F., Gennaro, C., Vadicamo, L., Vairo, C.: An image retrieval system for video. In: Similarity Search and Applications. SISAP 2019. pp. 332–339. Springer (2019)

8. Carrara, F., Vadicamo, L., Gennaro, C., Amato, G.: Approximate Nearest Neighbor Search on Standard Search Engines. In: Skopal, T., Falchi, F., Lokoč, J., Sapino, M.L., Bartolini, I., Patella, M. (eds.) Similarity Search and Applications. pp. 214–221. Springer International Publishing, Cham (2022)
9. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
11. Heller, S., Gsteiger, V., Bailer, W., Gurrin, C., Jónsson, B.Þ., Lokoč, J., Leibetseder, A., Mejzlík, F., Peška, L., Rossetto, L., 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)
12. Hezel, N., Schall, K., Jung, K., Barthel, K.U.: Efficient search and browsing of large-scale video collections with vibro. In: MultiMedia Modeling. pp. 487–492. Springer International Publishing, Cham (2022)
13. Lokoč, J., Mejzlík, F., Souček, T., Dokoupil, P., Peška, L.: Video search with context-aware ranker and relevance feedback. In: MultiMedia Modeling. pp. 505–510. Springer International Publishing, Cham (2022)
14. 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)
15. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5222–5229. IEEE (2021)
16. 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. arXiv preprint arXiv:2207.14757 (2022)
17. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. Tech. rep., Tech. Rep., Technical report, OpenAI (2022)
18. Revaud, J., Almazan, J., Rezende, R., de Souza, C.: Learning with average precision: Training image retrieval with a listwise loss. In: International Conference on Computer Vision. pp. 5106–5115. IEEE (2019)
19. Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., Vrochidis, S.: Interactive video retrieval in the age of deep learning - detailed evaluation of VBS 2019. IEEE Transactions on Multimedia pp. 1–1 (2020)
20. 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)
21. 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 - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023. Springer (2023)
22. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing **18**(7), 1512–1523 (2009)
23. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: VarifocalNet: An IoU-aware dense object detector. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2021)