VISIONE at Video Browser Showdown 2021

Giuseppe Amato[0000−0003−0171−4315], Paolo Bolettieri[0000−0002−5225−4278], Fabrizio Falchi[0000−0001−6258−5313], Claudio Gennaro[0000−0002−3715−149X], Nicola Messina[0000−0003−3011−2487], Lucia Vadicamo[0000−0001−7182−7038], and Claudio Vairo[0000−0003−2740−4331]

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

Abstract. This paper presents the second release of VISIONE, a tool for effective video search on large-scale collections. It allows users to search for videos using textual descriptions, keywords, occurrence of objects and their spatial relationships, occurrence of colors and their spatial relationships, and image similarity. One of the main features of our system is that it employs specially designed textual encodings for indexing and searching video content using the mature and scalable Apache Lucene full-text search engine.

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

1 Introduction

In the last decade, we have witnessed an exponential growth of multimedia content, mainly due to the pervasive use of cameras and social media. However, as visual data (e.g. video and images) are usually poorly annotated or not annotated at all, the use of scalable content-based retrieval systems and techniques for automatic visual analysis have become crucial to managing large visual archives. In this paper, we present a content-based video retrieval system, called VISIONE, which leverages various artificial intelligence techniques for automatic analysis of video keyframes in synergy with specially designed textual encoding of the visual content that facilitates the use of mature and scalable full-text search technologies for indexing and searching large-scale video collections.

A first release of VISIONE [1,6], which participated in the 2019 edition of the Video Browser Showdown (VBS) [11], is described in details in [2]. VBS is an international video search competition that is held annually since 2012 [13]. The V3C1 dataset [5], consisting of 7,475 videos, has been used in the competition since 2019. So far, three types of search tasks are considered in the competition: *Known-Item-Search (KIS)*, *textual KIS* and *Ad-hoc Video Search (AVS)*. The

J. Lokoč et al. (Eds.): MMM 2021, LNCS 12573, pp. 473-478, 2021.
Final authenticated pub.: https://doi.org/10.1007/978-3-030-67835-7_479

KIS task simulates the situation in which a user wants to find a particular video clip that he/she has watched before performing the search. The textual KIS concerns the case in which the user wants to find a particular video clip that he/she has never seen but of which a detailed textual description is provided. For the AVS task, instead, a general textual description is provided to the user who is asked to find as many video shots as possible that fit the given description.

One of the main limitations of the first version of VISIONE was the poor performance on textual KIS tasks. In facts, in our first participation in the VBS competition, the search in VISIONE was based only on object detection, colors, scene tags and visual similarity, and this proved to be not good enough to resolve textual KIS tasks in a reasonable time. To overcome this limitation, in this new release of our system, we integrated a retrieval module that allows searching for a target scene using natural language queries. Moreover, inspired by several systems that participated in previous editions of VBS, like [10,8,12], we introduced the possibility of performing a temporal search, where the user can describe two consecutive (or temporally close) keyframes of the same target video. Finally, several improvements have been made to the interface and in the selection of the best scoring functions used for ranking the results. All these novel aspects of our system are described in Section 3. The next section, instead, provides an overview of VISIONE and its functionalities.

2 VISIONE Video Search System

VISIONE provides several search functionalities in order to allows a user to search for a video by formulating textual or visual queries describing the content of a scene of a target video. In particular, it supports:

- *query by scene description*: the user can provide a textual description in natural language (e.g. "A tennis player serving a ball on the court");
- *query by keywords*: the user can specify keywords related to the target scene (e.g. "tennis, indoor, athlete, action");
- *query by object location*: the user can draw on a canvas simple diagrams to specify the objects that appear in a target scene and their spatial locations;
- *query by color location*: the user can specify some colors present in a target scene and their spatial locations;
- *query by visual example*: an image can be used as a query to retrieve video scenes that are visually similar to it. The image can be selected in the browsing interface as one of the results of a previous search iteration, or uploaded from URL/local file system.

Moreover, some filters are available to specify the aspect ratio of the target scene and if it is in color or in b/w. Figure 1 shows a screenshot of the search interface.

To support the above mentioned search functionalities, VISIONE exploits content analysis and artificial intelligence techniques to understand and represent the visual content of the video keyframes, including (i) a Transformer

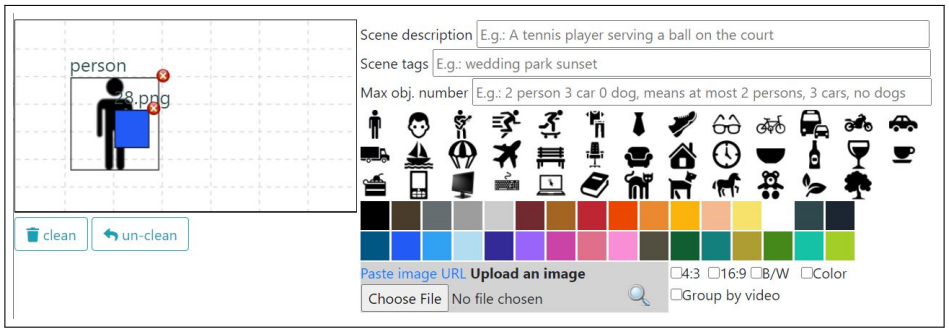

Fig. 1. Search Interface. (Color figure online)

Encoder Reasoning Network [9] to extract relation-aware textual and visual features that enable our system to search images using textual descriptions; (ii) an image annotation engine [4] to extract scene tags; (iii) state-of-the-art object detectors to identify and localize objects in the video keyframes; (iv) spatial colors histograms to identify dominant colors and their locations: (v) the R-MAC [14] deep visual descriptors to support the similarity search functionality.

One of the main peculiarity of our system is that all the different descriptors extracted from the video keyframes (features, scene tags, colors/object classes and locations) as well as the queries formulated by the user through the search interface (e.g., keywords describing the target scene and/or diagrams depicting objects and colors locations) are encoded using specifically-designed textual representations (see [2] for the details). This choice allows us to exploit mature and scalable full-text search technologies for indexing and searching large-scale video database without the need to implement dedicated access methods. In particular, VISIONE relies on the Apache Lucene full-text search engine.

3 New VISIONE Functionalities for VBS 2021

This section provides an overview of the improvements performed to the system compared to the first release of VISIONE that participated in VBS 2019.

Query by Textual Description To address the limitations of the previous version of VISIONE during the textual KIS, in this improved version we added an ad-hoc subsystem for searching keyframes using textual descriptions. Textual descriptions are full natural language sentences, usually between 5 to 50 words in length, describing a visual scene. For example, a valid textual description could be "*A tightly packed living room with a tv screen larger than the fireplace right beside it*". These textual descriptions can include objects details, expressed using their physical or semantic attributes, and they can specify the spatial or abstract relationships linking objects together.

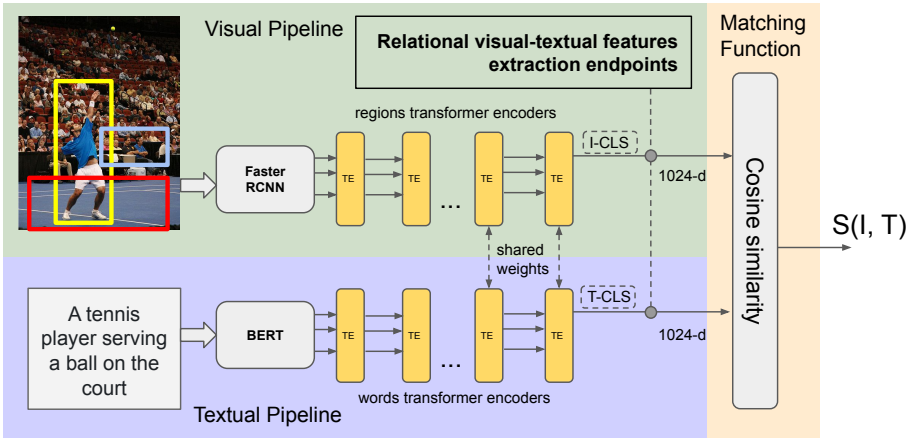

Fig. 2. Overview of the TERN architecture. Orange boxes are Transformer Encoder (TE) layers. Final TE layers share their weights for better stability during the training phase. (Color figure online)

This visual search using natural language descriptions as a query is achieved by using a recently developed deep neural network architecture, called Transformer Encoder Reasoning Network (TERN) [9], which is able to match images and sentences in a highly-semantic common space. The core of the architecture is constituted of recently introduced deep relational modules called *transformer encoders* [15], which can spot out hidden intra-object relationships. In particular, in the visual pipeline, a stack of transformer encoders try to find links between image regions pre-extracted using a state-of-the-art object detector (Faster-RCNN); in the textual pipeline, using a pretrained BERT model plus another stack of transformer encoder layers, the model searches for relationships between sentence words. An overview of the architecture is shown in Figure 2.

The extracted cross-modal features are normalized and in principle very similar to visual descriptors like RMAC [14]. Hence we indexed them using the same textual encoding that we already exploited to index the RMAC descriptors (see [2,3]).

Temporal Query To support temporal queries, in the new version of our system, we have added a second canvas and associated input text boxes to the user interface, that allows users to simultaneously search for two keyframes that are temporally close in a video segment but that are different in the represented content. The search is executed by performing two queries to the index, each providing its own output results. The resulting keyframes, which belong to the same video and whose temporal distance is less than a given threshold δ, are then combined as pairs and shown in the result section of the interface. We use $\delta = 20$ seconds, however we plan to integrate the possibility to specify a differ-

ent temporal threshold in the user interface. In this way, the user can exploit temporal relation between video keyframes when searching for a target video.

Improvements in the Searching Implementation The search process in VISIONE relies on five search operations which implement the five search functionalities presented in Section 2. The results of these search operations are combined and ranked according to some text scoring functions (see [2] for more details). In the first implementation of VISIONE, we selected the text scoring function to be used for each search operation by performing some very preliminary tests and by (subjectively) estimating the performance of the system. Recently, in [2], we performed a more in-depth and objective analysis to select the best rankers combination for our system. In particular, we tested 64 different implementations of our system using all the queries output collected during the participation at the VBS2019 challenge in order to select the configuration that has the best performance in terms of effectiveness (i.e. how good is the system in returning at least one relevant result in the top positions of the result set). We used this newly established configuration in the new release of the system.

User Interface Some improvements have also been made to the VISIONE user interface. We have integrated the possibility to search by similarity also using images uploaded from a URL or file system, as previously only images from the indexed data collection were allowed to be used as query examples. In addition, to boost the efficiency of our system during AVS tasks, we have added the possibility of selecting multiple images to be submitted as a response while automatically removing from the browsing interface all images that have already been submitted during the running AVS session.

4 Conclusion and Future Work

In this paper, we presented the second version of the VISIONE system, focusing on the new functionalities that we integrated in our system to better handle both KIS and AVS tasks. However, we plan to further improve our system in several ways, including exploiting video-text matching approaches (now the system uses only image-text matching), different color analysis techniques, more advanced techniques for organizing search results, and the use of textual speech and OCR annotations that are already provided by the VBS community. Moreover, we would like to integrate collaborative browsing and search functionalities. Finally, we are investigating the possibility of improving the bounding-box search tool by realizing a more precise match between user-defined rectangles during query and image bounding-boxes. The idea is to define a similarity function between two images based on the aggregation of the degree of overlap between the image bounding-boxes. Since there are several ways in which the bounding-boxes can be matched, the computation of this similarity defines an assignment problem, which can be solved in theory with the well-known Hungarian algorithm [7].

Given the complexity of the algorithm, this solution will presumably only be used to reorder the result-set of a query.

Acknowledgements

This work was partially funded by: AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911; AI4EU project (EC, H2020, n. 825619); AI4ChSites, CNR4C program (Tuscany POR FSE 2014-2020 CUP B15J19001040004). *The final authenticated version is available online at* `https://doi.org/10.1007/978-3-030-67835-7_47`

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 International Publishing (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. arXiv preprint arXiv:2008.02749 (2020)
3. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Vadicamo, L.: Large-scale instance-level image retrieval. Information Processing & Management p. 102100 (2019)
4. Amato, G., Falchi, F., Gennaro, C., Rabitti, F.: Searching and annotating 100M images with yfcc100m-hnfc6 and mi-file. In: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing. pp. 26:1–26:4. CBMI '17, ACM (2017)
5. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3C1 Dataset: An evaluation of content characteristics. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 334—-338. ICMR '19, Association for Computing Machinery (2019)
6. 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 International Publishing (2019)
7. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
8. Lokoč, J., Kovalčík, G., Souček, T.: VIRET at Video Browser Showdown 2020. In: MultiMedia Modeling. pp. 784–789. Springer International Publishing (2020)
9. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. In: International Conference on Pattern Recognition (ICPR) 2020 (Accepted) (2020)
10. Nguyen, P.A., Wu, J., Ngo, C.W., Francis, D., Huet, B.: Vireo @ video browser showdown 2020. In: MultiMedia Modeling. pp. 772–777. Springer International Publishing (2020)
11. 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)

12. Sauter, L., Amiri Parian, M., Gasser, R., Heller, S., Rossetto, L., Schuldt, H.: Combining boolean and multimedia retrieval in vitrivr for large-scale video search. In: MultiMedia Modeling. pp. 760–765. Springer International Publishing (2020)
13. Schoeffmann, K.: Video browser showdown 2012-2019: A review. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–4 (2019)
14. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. arXiv preprint arXiv:1511.05879 (2015)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)