VISIONE at Video Browser Showdown 2022

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
{giuseppe.amato,paolo.bolettieri,fabio.carrara,fabrizio.falchi,
claudio.gennaro,nicola.messina,lucia.vadicamo,claudio.vairo}@isti.cnr.it

Abstract. VISIONE is a content-based retrieval system that supports various search functionalities (text search, object/color-based search, semantic and visual similarity search, temporal search). It uses a full-text search engine as a search backend. In the latest version of our system, we modified the user interface, and we made some changes to the techniques used to analyze and search for videos.

Keywords: Content-based video retrieval · Video search · Information search and retrieval · Surrogate text representation

1 Introduction

In the last years, we witnessed an explosive growth in the amount of multimedia data present on the web. Nowadays, the pervasive use of social networks and cameras for surveillance applications generates a tremendous amount of multimedia content, which must be indexed for being efficiently browsed and retrieved. Visual data, in the past, were indexed using manual annotations. In high-data regimes, manual labeling is not feasible, so many techniques have been developed to search images or videos based on their content. In this paper, we describe VISIONE, a content-based video retrieval system for efficient and effective video search, which employs state-of-the-art deep learning techniques to extract visual content from videos at different levels of abstraction—from colors to high-level semantics. It employs a special textual encoding of the visual features and off-the-shelf full-text search tools for indexing all the different descriptors. With this expedient, on the one hand, we improve storage utilization by using the same data structure to store diverse multi-modal data; on the other, we exploit the scalability of off-the-shelf text search engines.

In this paper, we aim at describing the latest version of VISIONE for participating to the Video Browser Showdown (VBS) [10,17]. The first version of the tool [1,2] and the second [3] participated in previous editions of the competition, VBS 2019 and VBS 2021, respectively. VBS is an international video search competition that is held annually since 2012 and comprises three tasks, consisting

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13142, pp. 543–548, 2022.
https://doi.org/10.1007/978-3-030-98355-0_52

of visual and textual known-item search (KIS) and ad-hoc video search (AVS) [10,17]. Starting with VBS 2022, the V3C1 dataset [6] will be extended by V3C2 [18], obtaining a grand total of 17,235 video files and 2,300 h of video content. It is clear that the competition is becoming every year more and more challenging.

By analyzing the issues and the failure cases of VISIONE at previous VBS editions, we improved the tool in several ways, as described in the next section.

2 System Overview

VISIONE integrates the following search functionalities:

- **spatial object-based search**: the user can draw simple bounding-boxes on a canvas to specify the objects (along with their spatial locations) appearing in a target video scene.
- **spatial color-based search**: similar to object-based search, the user can draw simple bounding-boxes to specify the colors (along with their spatial locations) that appear in a target scene.
- **free text search**: the user can provide a textual description, in natural language, of a video scene.
- **visual similarity search**: the user can use an image (selected from the results of a previous search or from the web) as a query to search for video keyframes *visually* similar to it.
- **semantic similarity search**: the user can select an image from a search results list to retrieve video keyframes that are *semantically* similar to it.
- **temporal search**: the above search features can be used to simultaneously search for two different scenes that are temporally close in a video clip.

One of the main characteristics of our system is that all the features extracted from the video keyframes, as well as from the user query, are transformed into textual encodings so that an off-the-shelf full-text search engine is employed to support large-scale indexing and searching (see [3] for further details).

While the object/color and similarity search functionalities have been present in VISIONE since its first version [1], the text search and the temporal search were introduced last year [3]. The semantic similarity search (Sect. 2.1), although not initially foreseen in the second release of VISIONE [3], was included in the system as a test functionality a few weeks before the last VBS competition (June 2021). Since our team users gave positive feedback on this new feature, we decided to integrate it into this year's system as well. In addition, some techniques used for extracting dominant colors, objects, and visual features have been modified since just before the last VBS competition, as described in Sect. 2.2. We would like to note that VISIONE used to support a *keyword search* (query by scene tags) that has been removed from the system this year for two main reasons: on the one hand to improve the usability of the system (having too many search options could be confusing especially for novice users); on the other because we noticed that during the last competition this tool was rarely used in favor of the textual search that gives the possibility to use more detailed descriptions of a target scene than a list of tags. Finally, looking forward to VBS 2022, the user interface has been redesigned as described in Sect. 2.3.

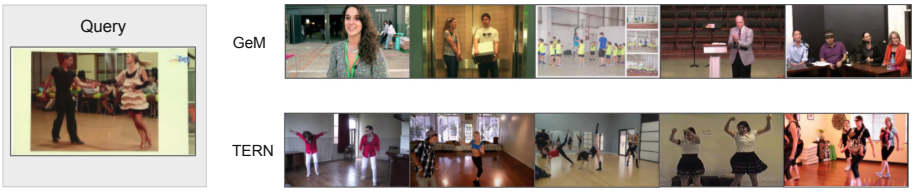

Fig. 1. Qualitative results on GeM and TERN features for image-to-image retrieval on a sample image query.

2.1 Semantic Image Retrieval

In the last version of VISIONE, we employed multi-modal features extracted using a recently proposed architecture called TERN (Transformer Encoder Reasoning Network) to support the text-to-visual retrieval [11,12]. We recently observed that the visual features—the ones extracted from the visual path of TERN—carry very-high level semantics and are therefore good candidate descriptors for performing *semantic* content-based image retrieval (S-CBIR). This is reasonable, as the multi-modal contrastive learning used in TERN generates a space in which images with similar textual descriptions are close together.

The mainstream CBIR research is developed and evaluated in instance-retrieval scenarios, which do not require a deep semantic understanding to be solved. This is the case for the R-MAC [8] or GeM [16] descriptors. They cannot embed high-level semantics and entity-entity relationships, as they are more sensible to simple object classes and to very specific low-level patterns. Differently, the supervision with textual embeddings in TERN have the side effect of creating highly-semantic visual features, which can account for actions, object attributes, and relationships between multiple actors in the scene. We qualitatively compare the GeM versus the TERN features for a simple query image in Fig. 1. As we can notice, the GeM features can retrieve low-level matching images, like images with persons or with similar patterns in the background. Instead, the TERN descriptors can correctly capture the concept of *dancing people*.

Having very different natures, in VISIONE we incorporated both the GeM and TERN descriptors, covering both the instance and semantic retrieval needs.

2.2 Changes to Some Features Used in VISIONE

Compared to the last system description [3], we have modified and/or integrated some features used by our system for object, color, similarity, and text search. We used VarifocalNet [21], which is a dense object detector, instead of YOLOv3 [15], and we replaced R-MAC features [8], previously used to assess visual similarity, with features extracted using GeM [16]. Moreover, we significantly revised the color palette and the extraction of colors used for our color-based search. Previously, as indicated in [2,3], we used a color palette consisting of 32 colors, and we classified the color of each image pixel using a k-nearest neighbor

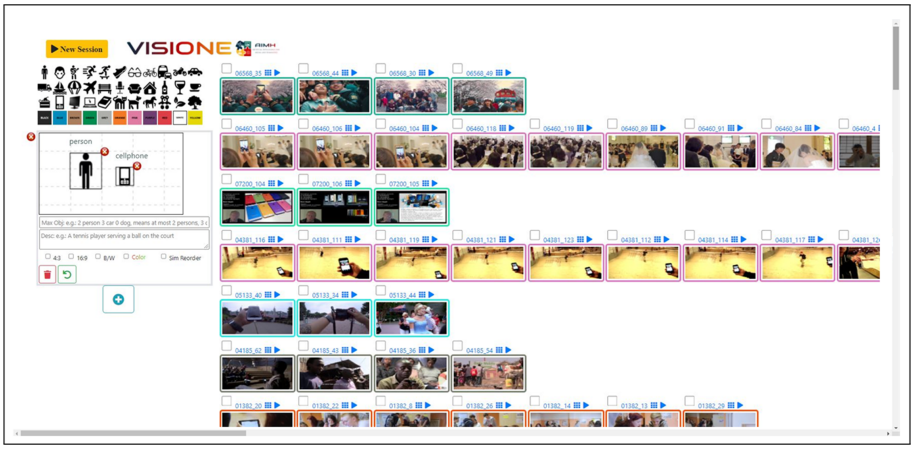

Fig. 2. New user interface. (Color figure online)

search between the actual pixel color and the colors in our palette. Nevertheless, many studies in the field of anthropology, visual psychology, and linguistics pointed out that some colors appear more memorable than others and that some basic color terms are used consistently and with consensus in different languages [5,7,19]. In particular, a set of 11 basic color terms (*white, black, red, green, yellow, blue, brown, purple, pink, orange, and gray*) are often used as universal color categories. Thus we decided to restrict our palette to these basic colors. To this purpose, we employed two *chip-based color naming* approaches [4,20] that, through the use of Probabilistic Latent Semantic Analysis and a parametric fuzzy model, respectively, provide ready-to-use hash tables to map RGB values to color names. We use both models to assign color terms to each pixel and then calculate the dominant colors for each of the 7×7 image cells. Finally, taking inspiration from [13] we are deploying in our system the CLIP features [14], aggregated through time, for text-to-video retrieval. In fact, with the increased dataset size, we need to better understand fine-grained actions. This aims to increase the discriminative power of the system—and in turn enhance the performance on textual-KIS and AVS tasks—by using state-of-the-art cross-modal features and leveraging the temporal domain.

2.3 New User Interface

From the analysis of the system logs collected in the last VBS, we realized that in some KIS tasks we were not able to submit the correct answer even though there was a keyframe of the target video (but not of the right shot) in the first positions of our result list. Therefore, we modified our user interface to improve the visualization and browsing of the results. In particular, inspired by other systems participating at VBS (e.g., [9,13]) we decided to place the search interface side by side with the browsing interface (respectively on the left and

right of the GUI) and to display the search results by grouping those from the same video on a single row (Fig. 2). Now, the results from the same video can be inspected by moving horizontally in the browsing interface, those of different videos by moving vertically. In comparison to the previous interface—limited to only a total of 600 keyframes not grouped by video—this change allows us to set a maximum number of results for each video, enabling the browsing of a greater number of results. In addition, other minor adjustments have been made in the GUI, such as the possibility to add two or more canvas and text boxes for temporal queries (in the old interface the double canvas was fixed) and to display some details or select some search options when hovering the mouse over a result.

3 Conclusions and Future Work

In this paper, we presented the latest version of the VISIONE system for participating at the next edition of VBS. In particular, to improve the results visualization and browsing, we redesigned the interface taking inspiration from other VBS systems. We used the already deployed TERN features to perform semantic CBIR, and we replaced some color, similarity, and text search features to align with the current state-of-the-art in image and video retrieval. We plan to further improve the system by mixing TERN and CLIP features to obtain highly-semantic features for text-to-video retrieval. Furthermore, we plan to develop a system to manually weight different portions of the text to have finer control over the text queries.

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).

References

1. Amato, G., et al.: VISIONE at VBS2019. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., Vrochidis, S. (eds.) MMM 2019. LNCS, vol. 11296, pp. 591–596. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05716-9_51
2. Amato, G., et al.: The VISIONE video search system: exploiting off-the-shelf text search engines for large-scale video retrieval. J. Imaging **7**(5), 76 (2021)
3. Amato, G., et al.: VISIONE at video browser showdown 2021. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 473–478. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_47
4. Benavente, R., Vanrell, M., Baldrich, R.: Parametric fuzzy sets for automatic color naming. JOSA A **25**(10), 2582–2593 (2008)
5. Berlin, B., Kay, P.: Basic Color Terms: Their Universality and Evolution. University of California Press, Berkeley (1991)

6. 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. Association for Computing Machinery (2019)
7. Boynton, R.M., Olson, C.X.: Salience of chromatic basic color terms confirmed by three measures. Vision. Res. **30**(9), 1311–1317 (1990)
8. Gordo, A., Almazan, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. Int. J. Comput. Vision **124**(2), 237–254 (2017)
9. Heller, S., et al.: Towards explainable interactive multi-modal video retrieval with vitrivr. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 435–440. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_41
10. Lokoč, J., et al.: Is the reign of interactive search eternal? Findings from the video browser showdown 2020. ACM Trans. Multimed. Comput. Commun. Appl. **17**(3), 1–26 (2021)
11. Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S.: Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. arXiv preprint arXiv:2008.05231 (2020)
12. 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)
13. Peška, L., Kovalčík, G., Souček, T., Škrhák, V., Lokoč, J.: W2VV++ BERT model at VBS 2021. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 467–472. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_46
14. Radford, A., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
15. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. CoRR abs/1804.02767 (2018)
16. 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)
17. Rossetto, L., et al.: Interactive video retrieval in the age of deep learning - detailed evaluation of VBS 2019. IEEE Trans. Multimedia **23**, 243–256 (2020)
18. Rossetto, L., Schoeffmann, K., Bernstein, A.: Insights on the V3C2 dataset. arXiv preprint arXiv:2105.01475 (2021)
19. Sturges, J., Whitfield, T.A.: Salient features of munsell colour space as a function of monolexemic naming and response latencies. Vision. Res. **37**(3), 307–313 (1997)
20. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Trans. Image Process. **18**(7), 1512–1523 (2009)
21. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: VarifocalNet: an IoU-aware dense object detector. In: Conference on Computer Vision and Pattern Recognition, pp. 8514–8523. IEEE, June 2021