VISIONE at VBS2019

Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole,
Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo

Institute of Information Science and Technologies (ISTI), Italian National Research Council (CNR), Via G. Moruzzi 1, 56124 Pisa, Italy
name.surname@isti.cnr.it

Abstract. This paper presents VISIONE, a tool for large–scale video search. The tool can be used for both known-item and ad-hoc video search tasks since it integrates several content-based analysis and retrieval modules, including a keyword search, a spatial object-based search, and a visual similarity search. Our implementation is based on state-of-the-art deep learning approaches for the content analysis and leverages highly efficient indexing techniques to ensure scalability. Specifically, we encode all the visual and textual descriptors extracted from the videos into (surrogate) textual representations that are then efficiently indexed and searched using an off-the-shelf text search engine.

Keywords: Content-based video retrieval · Video search · Known Item Search · Convolutional Neural Networks

1 Introduction

The Video Browser Showdown (VBS) [4,10] is an international video search competition that evaluates the performance of interactive video retrievals systems. It is performed annually since 2012; however, it is becoming increasingly challenging as the used video archive grows and new query tasks are introduced in the competition. The VBS 2019 uses the V3C1 dataset that consists of 7,475 video files (amounting for 1000h of video content) and encompasses three content search tasks: *Known-Item-Search (KIS)*, *textual KIS* and *Ad-hoc Video Search (AVS)*. The KIS task models the situation in which someone wants to find a particular video clip that he has already seen, assuming that it is contained in a specific collection of data. The textual KIS is a variation of the KIS task, where the target video clip is no longer visually presented to the participants of the challenge but it is rather described in details by text. This task simulates situations in which a user wants to find a particular video clip, without having seen it before, but knowing the content of the target video exactly. For the AVS task, instead, a general textual description is provided (e.g. "A person playing guitar outdoors") and participants need to find as many correct examples as possible, i.e. video shots that fit the given description.

In this paper, we present the first version of VISIONE, a system which integrates several search capabilities for efficient video retrieval. Specifically, it supports:

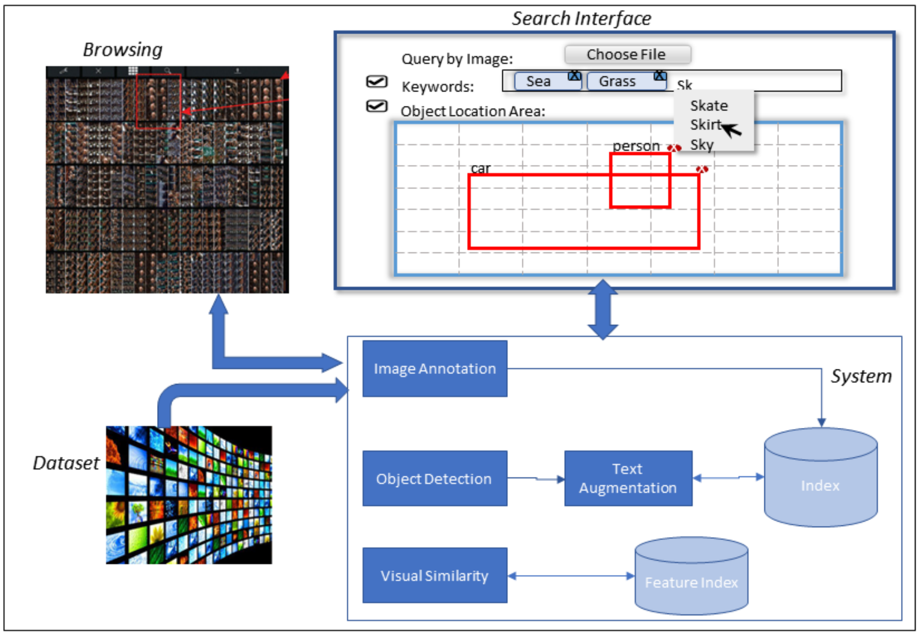
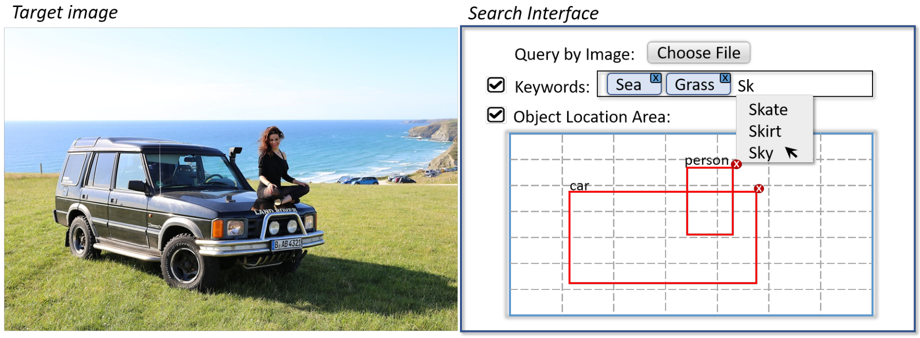


Fig. 1. A sketch of VISIONE's architecture.

- query by keywords, i.e. the user can specify some textual keywords/tags related to the target video segment(s) to be retrieved;
- query by object location, i.e. the user can draw simple object-based diagrams specifying the spatial location of the objects appearing in the target scene;
- query by example, i.e. the user can upload an image or select a keyframe of a video to search for videos with similar visual content.

State-of-the-art deep learning approaches are used to analyze and annotate about 1.1M of representative keyframes selected from V3C1 dataset. We encode all the features extracted from the videos (visual features, tags, object locations, and metadata) into textual representations that are then indexed using inverted files. We use a text surrogate representation [6], which has been appositely extended to support efficient spatial object queries on large scale dataset. It will be possible to build queries by placing wanted objects in the scene and to efficiently search for compatible images in an interactive way. This choice allows us to exploit efficient and scalable search technologies and platform used nowadays for text retrieval. In particular, VISIONE relies on the Elasticsearch [1].

[1] https://github.com/elastic/elasticsearch



Query by Image: Choose File

☑ Keywords: Sea Grass Sk

☑ Object Location Area:

Fig. 2. VISIONE user interface for: keyword-based, object spatial position and the query by image search.

2 VISIONE: Content Based Video Retrieval System

VISIONE relies heavily on deep learning techniques, trying to bridge the semantic gap between text and image. In our system, we focus on bridging these two descriptions using the following approaches:

- keyword search: we use keywords including scenes, places or concepts (e.g. outdoor, building, sport) to query video shots. Our image annotation system is based on different Convolutional Neural Networks to extract scene attributes.
- object location search: we use the location of the objects in a scene to query video shots. We exploit the efficiency of YOLO [2] as a state of the art real-time object detection system to retrieve the video shot containing the objects sketched by the user.
- visual similarity search: we use an image as query to retrieve the most similar video shots. The similarity search is performed by calculating the cosine similarity of the visual features on the entire visual data set, represented using the R-MAC [16] visual descriptor.

2.1 Keyword-based Search

Since, the categories with which Hybrid-CNN was trained are insufficient to associate relevant tags to the images, VISIONE exploits our automatic annotation system to annotate untagged images [1]. This system is based on YFCC100M-HNfc6, a set of deep features we extracted from the YFCC100M dataset [15]. The YFCC100M-HNfc6 feature dataset was created using the Caffe framework [8]. In particular, we use the neural network Hybrid-CNN whose model and weights are publicly available in the Caffe Model Zoo. The Hybrid-CNN[3] was trained

[2] https://pjreddie.com/darknet/yolo/

[3] http://github.com/BVLC/caffe/wiki/Model-Zoo.

on 1,183 categories: 205 scene categories from Places Database (Places205) and 978 object categories from the train data of ILSVRC2012 (ImageNet)[14]. Our image annotation system is based on an unsupervised approach to extract the implicitly existing knowledge in the huge collection of unstructured texts describing the images of YFCC100M dataset, allowing us to label the images without using a training model. The image annotation system also exploits the metadata of the images validated using WordNet [5]. For the competition the idea is to integrate the new Place Dataset (Places365-CNN [18]), other concepts as 345 TRECVID SIN concepts [3], the 239 categories from Fudan-Columbia Video Dataset (FCVID) [9].

2.2 Object Location Search

Following the idea that the human eye is able to identify objects in the image very quickly, we decide to take advantage of the new technologies available to search for object instances in order to retrieve the exact video shot.
For this purpose, we use YOLOv3 [13] as object detector, both because it is extremely fast and because of its accuracy. Our image query interface is subdivided into a $N \times N$ grid in the same way that YOLO segments images to detect objects. Each object detected in the single image I by YOLO is indexed using a specific encoding ENC conceived to put together the location and the class corresponding to the object ($cod_{pos} cod_{text}$). The idea of using YOLO to detect objects within video has already been exploited in VBS, e.g. by Truong et al. [17], but our approach is distinguished by being able to encode the class and the location of the objects in a single textual description of the image, allowing us to search by similarity using a standard text search engine. Basically for each I entry on the index, we have a space-separated concatenation of ENCs, one for all the possible cells (cod_{pos}) in the grid that contains the object (cod_{text}). For example, for the image in Figure 2 the object *car* is indexed with the sequence $c_1 car$ $c_2 car$ ... $c_n car$, where c_i is the code of the i-th cell containing the car.

In order to approach the KIS task, the user can take advantage of our UI to sketch the objects appearing in the target video by specifying the desired location for each object (as shown in the “Object Location Area” of Figure 2).

2.3 Visual Similarity Search

VISIONE also supports content-based visual search functionalities, i.e., it allows users to retrieve scenes containing keyframes visually similar to a query image given by example. To start the search the user can select any keyframe of a video as query (e.g. any one presented in the results-set of a previous search) or he can directly upload an image (e.g. selected using an external image search engine, like Bing or Google). In order to represent and compare the visual content of the images, we use the Regional Maximum Activations of Convolutions (R-MAC) [16], which is a state-of-art descriptor for image retrieval. The R-MAC descriptor effectively aggregates several local convolutional features (extracted at multiple position and scales) into a dense and compact global image representation. We use the ResNet-101 trained model provided by Gordo et al. [7] as an R-MAC feature extractor since it achieved the best performance on standard benchmarks. To efficiently index the R-MAC descriptor we transform the deep features into a textual encoding suitable for being indexed by a standard full-text search engine, such as Elasticsearch. The process to transform the deep features into a textual representation is done as follows. We first use the Deep Permutation technique [2] to encode the deep features into a permutation vector, which is then transformed into a Surrogate Text Representation (STR) as described in [6]. The main rationale of this approach is that if two features are very close one to the other, they will have similar Deep Permutations and thus the corresponding text representations will be close as well. The advantage of using the textual encodings is that we can efficiently exploit off-the-shelf text search engines for performing image searches on large scale.

3 Conclusion

We present the first version of VISIONE, a system that will be used in the Video Browser Showdown 2019 challenge. The system can be used for both the Known Item Search or Ad-hoc Video Search tasks. It supports three types of queries: query by keyword, query by object location, and query by visual similarity. All the features used to represent the video keyframes are extracted using state-of-the-art deep learning approaches. VISIONE also exploits ad-hoc surrogate text encodings of the extracted features in order to use efficient technologies and platforms for text retrieval, without the need for the definition of dedicated access methods. Inspired by the SIRET system [11,12] that won the VBS2018, we plan to integrate other useful search capabilities (such as query-by-color sketches) in a future version of our system.

Acknowledgements

This work was partially funded by “Smart News: Social sensing for breaking news”, CUP CIPE D58C15000270008, by VISECH ARCO-CNR, CUP B56J17001330004, and by “Automatic Data and documents Analysis to enhance human-based processes” (ADA), CUP CIPE D55F17000290009. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. 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, CBMI 2017, Florence, Italy, June 19-21, 2017. pp. 26:1–26:4 (2017). https://doi.org/10.1145/3095713.3095740
2. Amato, G., Falchi, F., Gennaro, C., Vadicamo, L.: Deep permutations: Deep convolutional neural networks and permutation-based indexing. In: International Conference on Similarity Search and Applications. pp. 93–106. Springer (2016)

3. Awad, G., Snoek, C.G.M., Smeaton, A.F., Quénot, G.: Trecvid semantic indexing of video : A 6-year retrospective (2016)
4. Cobârzan, C., Schoeffmann, K., Bailer, W., Hürst, W., Blažek, A., Lokoč, J., Vrochidis, S., Barthel, K.U., Rossetto, L.: Interactive video search tools: a detailed analysis of the video browser showdown 2015. Multimedia Tools and Applications **76**(4), 5539–5571 (Feb 2017). https://doi.org/10.1007/s11042-016-3661-2
5. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
6. Gennaro, C., Amato, G., Bolettieri, P., Savino, P.: An approach to content-based image retrieval based on the lucene search engine library. In: International Conference on Theory and Practice of Digital Libraries, pp. 55–66 (2010)
7. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: End-to-end learning of deep visual representations for image retrieval. International Journal of Computer Vision **124**(2), 237–254 (2017)
8. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
9. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 352–364 (2018). https://doi.org/10.1109/TPAMI.2017.2670560
10. Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Awad, G.: On influential trends in interactive video retrieval: Video browser showdown 2015-2017. IEEE Transactions on Multimedia pp. 1–1 (2018). https://doi.org/10.1109/TMM.2018.2830110
11. Lokoč, J., Kovalčík, G., Souček, T.: Revisiting siret video retrieval tool. In: Schoeffmann, K., Chalidabhongse, T.H., Ngo, C.W., Aramvith, S., O'Connor, N.E., Ho, Y.S., Gabbouj, M., Elgammal, A. (eds.) MultiMedia Modeling. pp. 419–424. Springer International Publishing, Cham (2018)
12. Lokoč, J., Souček, T., Kovalčik, G.: Using an interactive video retrieval tool for lifelog data. In: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge. pp. 15–19. LSC '18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3210539.3210543
13. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv (2018)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
15. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Commun. ACM **59**(2), 64–73 (Jan 2016). https://doi.org/10.1145/2812802
16. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015)
17. Truong, T.D., Nguyen, V.T., Tran, M.T., Trieu, T.V., Do, T., Ngo, T.D., Le, D.D.: Video search based on semantic extraction and locally regional object proposal. In: Schoeffmann, K., Chalidabhongse, T.H., Ngo, C.W., Aramvith, S., O'Connor, N.E., Ho, Y.S., Gabbouj, M., Elgammal, A. (eds.) MultiMedia Modeling. pp. 451–456. Springer International Publishing, Cham (2018)
18. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)