VERGE in VBS 2023

Nick Pantelidis, Stelios Andreadis(✉), Maria Pegia, Anastasia Moumtzidou, Damianos Galanopoulos, Konstantinos Apostolidis, Despoina Touska, Konstantinos Gkountakos, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris

Information Technologies Institute/Centre for Research and Technology Hellas, Thessaloniki, Greece
{pantelidisnikos,andreadisst,mpegia,moumtzid,dgalanop,kapost, destousok,gountakos,heliasgj,stefanos,bmezaris,ikom}@iti.gr

Abstract. This paper describes VERGE, an interactive video retrieval system for browsing a collection of images from videos and searching for specific content. The system utilizes many retrieval techniques as well as fusion and reranking capabilities. A Web Application is also part of VERGE, where a user can create queries, view the top results and submit the appropriate data, all in a user-friendly way.

1 Introduction

VERGE is an interactive video retrieval system that provides various search capabilities inside a set of images, along with a Web Application for creating and running queries on the set, viewing the top results and submitting the appropriate ones. Over the past few years VERGE has participated many times in the Video Browser Showdown (VBS) competition [9] trying each time to better adapt to the competition's "Ad-Hoc Video Search" (AVS) and "Known Item Search - Visual/Textual" (KIS-V, KIS-T) tasks. This year the various search modules have been further improved and the Web Application has been updated in order to be even more user-friendly and fast-to-use.

The paper is structured as follows: Sect. 2 describes the overall framework of the system, Sect. 3 continues with a detailed description of the various retrieval modalities, Sect. 4 presents the user interface (UI) and its features, and the paper wraps up with Sect. 5 that briefly describes the future work.

2 The VERGE Framework

As shown in Fig. 1, the VERGE framework is composed of three layers. The first layer contains all the retrieval modalities that are applied on the datasets, i.e. V3C1, V3C2 [18] and the Marine Video Dataset. The outcomes are stored in a database (except for the ones from Text to Video Matching module that only runs on-the-fly). The second layer consists of the various services that accept queries and return the top results. The third layer is the Web Application that allows users to formulate and send queries, connects to the corresponding services and displays the results.

D.-T. Dang-Nguyen et al. (Eds.): MMM 2023, LNCS 13833, pp. 658–664, 2023.
https://doi.org/10.1007/978-3-031-27077-2_55

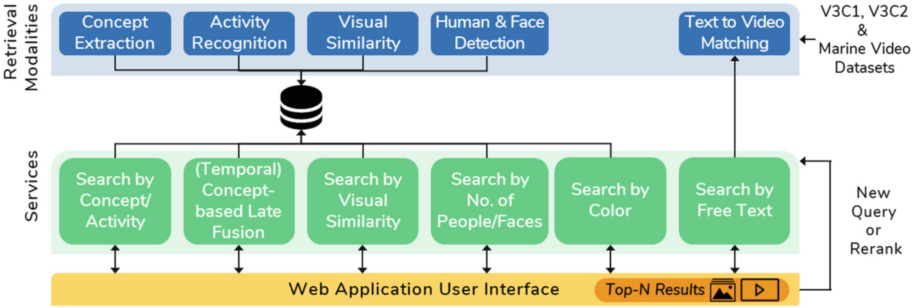


Fig. 1. The VERGE framework

3 Retrieval Modalities

3.1 Concept-Based Retrieval

This module annotates each keyframe of a shot with labels from a pool of concepts, which comprises 1000 ImageNet concepts, a selection of 298 concepts of the TRECVID SIN task [14], 500 event-related concepts, 365 scene classification concepts, 580 object labels as well as 22 sports classification labels. To obtain the annotation scores for the 1000 ImageNet concepts, we used an ensemble method, averaging the concept scores from two pre-trained models that employ different DCNN architectures, i.e. the EfficientNet-B4 [20] and InceptionResNetV2. To obtain scores for the subset from the TRECVID SIN task, we trained and employed a model based on the EfficientNet-B4 architecture on the official SIN task dataset. For the event-related concepts, we used the pre-trained model of EventNet [7]. Regarding the extraction of the scene-related concepts, we utilized the publicly available VGG16 model, fine-tuned on the Places365 dataset. Object detection scores were extracted using models pre-trained on the MS COCO and Open Images V4 datasets, with 80 and 500 detectable objects, respectively. To label sports in video frames, we constructed a custom dataset with Web images from sports and utilized it to train a model of the EfficientNetB3 architecture. Finally, to offer a cleaner representation of the concept-based annotations we employed the sentence-BERT [17] text encoding framework, to measure the text similarity between all concepts' labels. After inspecting the results, we manually formed groups of very similar concepts for which we created a common label and assign the max score of its members.

3.2 Spatio-temporal Activity Recognition

The activity detection and recognition module extracts human-related activities for each shot to enrich the filtering functionalities using the labels of the activities. A list of 400 pre-defined human related-activities and the corresponding scores were extracted for each shot using a 3D-CNN architecture. Especially, the configuration of the 3D-ResNet architecture with 152 layers according to [8] was used and the model's pre-trained weights were learned using the Kinetics-400 dataset [3]. During inference, the input shots fed to the model were pre-processed to be descended to the model's input dimension equal to 16 × 112 × 112 × 3.

3.3 Visual Similarity Search

The visual similarity search module uses as input the visual features of each shot and retrieves the most similar content using DCNNs. These features are the output of the last pooling layer of the fine-tuned GoogleNet architecture [15] and are used for globally representing the images. In order to allow fast and efficient indexing, an IVFADC index database vector is developed with these vectors [10].

3.4 Human and Face Detection

The human and face detection module aims to detect and count humans and human faces in each keyframe of each shot, so that the user can easily distinguish the results of single-human or multi-human activities. The detection of both human silhouettes (bodies) and human faces (heads) were extracted using a DCNN architecture, YoloV4 [1]. The model's initial weights are learned using the MS COCO [12] dataset and fine-tuned using the CrowdHuman dataset [19] that consists of crowd-center scenes where partial occlusions among humans or between humans and objects are possible to occur. During inference, the total number of humans and human heads is calculated and counted only in case the detected bounding box area is larger than a predefined threshold.

3.5 Text to Video Matching Module

The text-to-video matching module inputs a complex free-text query and retrieves the most relevant video shots from a large set of available video shots. We utilize the T×V model, a cross-modal video retrieval method presented in [6]. The T×V model utilizes multiple textual and visual features along with multiple textual encoders to build multiple cross-modal common latent spaces. The network is trained using video-caption pairs and learns to transform these pairs into multiple common latent spaces. The straightforward comparison between a video and a caption is possible in every individual space. Using a multi-loss-based training approach our network learns the overall similarity by optimizing the individual similarities. Regarding the training, a combination of four large-scale video caption datasets is used (e.g. MSR-VTTT [22], TGIF [11], ActivityNet [2] and Vatex [21]), and the improved marginal ranking loss [4] is used to train the entire network. As initial video shot representation, we utilize three different trained networks, i) the ResNet-152 deep network trained on the ImageNet-11k dataset, ii) the ResNeXt-101 network, pre-trained by weakly supervised learning on web images followed and fine-tuned on ImageNet [13], and iii) the CLIP model (ViT-B/32) [16]. As textual encoders, the networks utilize i) a feedforward encoder utilizing the CLIP-based generated embeddings, and ii) the textual sub-network (ATT) presented in [5].

3.6 Concept-Based Late Fusion

The concept-based late fusion module returns a list of shots, where each shot contains all the queried concepts. Specifically, the method uses two or more visual concepts (Sect. 3.1) and produces a sorted shots' list via a late fusion method. First, for each concept an independent list of shot probabilities $P = \{p_i\}_{i=1}^{i=n}$ is developed. The intersection of the concepts at shot layer is calculated and is sorted using the objective function

$$f(P) = \sum_{i=1}^{|P|} e^{-p_i} + \sum_{i,j=1,i \neq j}^{|P|} e^{-|p_i - p_j|}. \tag{1}$$

This function follows the assumption that the higher the concept probabilities or the more relevant the shots are, the higher their scores are. We deal with it using the difference of the probabilities for all concept pairs' combinations followed by the inverse exponential function to them.

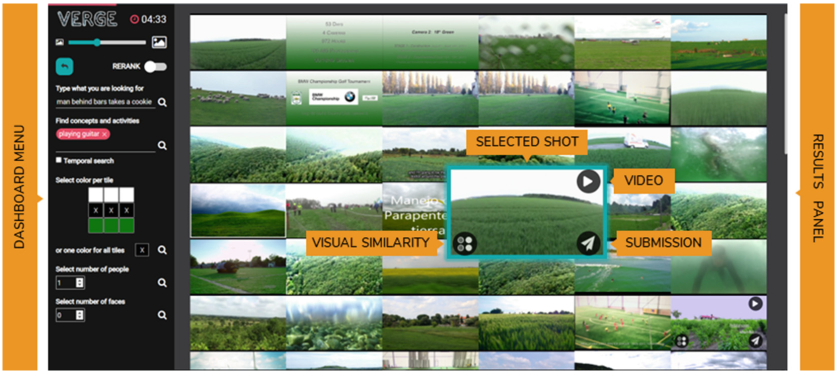
3.7 Temporal Late Fusion

The temporal late fusion module returns a list of tuples of shots, where each element of the list corresponds to the same video and contains all the queried concepts with respect to the given order. In particular, the module incorporates two or more visual concepts (Sect. 3.1) and produces a sorted list without duplicates via a late fusion method. At first, a list of shot probabilities is produced for each concept. Next, the intersection of concepts at video layer is calculated and the first tuple of each video, which respects the ordering of the concepts, is kept. The shots are sorted using an objective function that respects the same assumptions identified in the concept-based late fusion method (Sect. 3.6).

4 User Interface and Interaction Modes

The VERGE UI (Fig. 2) is a Web application that allows a user to easily create and run queries on the dataset and view the top results using the modalities that were described in the previous sections. They can also watch the corresponding video and submit the appropriate data during the VBS competition. The goal of the VERGE UI is to provide a user-friendly, compact, effective and fast tool for searching in image collections.

The UI has two main parts: the menu on the left and the results panel on the right. On the top of the *menu* there is a timer that counts down the remaining time for submission during a VBS task. Below there is a slider where a user can define the size of the images on the results panel, an undo button for restoring previous results, a rerank button for reranking the current results based on the next query, and then follow the various search modules. The first module is the free text search (Sect. 3.5) where the user can type anything in the form of free text and the second one allows the user to search from a list of pre-extracted



Type what you are looking for
man behind bars takes a cookie

Find concepts and activities
playing guitar ×

□ Temporal search

Select color per tile

or one color for all tiles

Select number of people
1

Select number of faces
0

Fig. 2. The VERGE user interface (Color figure online)

concepts and activities (Sects. 3.1, 3.2). Multiple selection is also supported for late fusion (Sect. 3.6) as well as temporal fusion (Sect. 3.7), if the corresponding checkbox is checked. Search by the color of the image is possible by coloring a 3 × 3 grid. Finally, there are search options based on the number of people or faces (Sect. 3.4) visible in an image.

The *results panel* contains the top results in a grid view. When an image is clicked, a pop-up panel appears that shows all the available shots of the corresponding video. Hovering over an image, three buttons appear. One on the bottom-left corner of the image that, when clicked, returns visually similar images (Sect. 3.3), one on the bottom-right corner for submitting this shot and one on the top-right corner that plays the respective video. Under the video player there is a button to submit directly the time of the video.

To demonstrate the features of VERGE, we shortly describe three use cases. For an AVS task that asks for shots of a single person playing guitar, the user can select the concept "playing guitar" and rerank the results by selecting only one person to appear. For a KIS-V query that searches for a video that shows a grassland on the bottom and the white sky on the top, the user can utilize the search by color, painting the first row white and the third row green (Fig. 2). Lastly, for a KIS-T query that asks for a video that shows "a man behind bars taking a cookie from a tray held" the exact words can be used per se in the free-text search.

5 Future Work

Every year we try to improve the retrieval performance by increasing the response time and the effectiveness of the algorithms, as well as to make the VERGE UI more intuitive, user-friendly and fast. The whole system will be evaluated in the VBS 2023 competition and the participating experience will drive the future steps, regarding both the search methodologies and the UI.

Acknowledgements. This work has been supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-101004152 CALLISTO, H2020-833464 CREST, H2020-101070250 XRECO, and H2020 - 101021866 CRiTERIA.

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
2. Caba Heilbron, F., et al.: ActivityNet: a large-scale video benchmark for human activity understanding. In: Proceedings of IEEE CVPR 2015, pp. 961–970 (2015)
3. Carreira, J., Zisserman, A.: Quo Vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
4. Faghri, F., Fleet, D.J., et al.: VSE++: improving visual-semantic embeddings with hard negatives. In: Proceedings of BMVC 2018 (2018)
5. Galanopoulos, D., Mezaris, V.: Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In: Proceedings of ACM ICMR 2020 (2020)
6. Galanopoulos, D., Mezaris, V.: Are all combinations equal? Combining textual and visual features with multiple space learning for text-based video retrieval. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13804, pp. 627–643. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25069-9_40
7. Guangnan, Y., Yitong, L., et al.: Eventnet: a large scale structured concept library for complex event detection in video. In: Proceedings of ACM MM 2015 (2015)
8. Hara, K., et al.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of IEEE CVPR 2018 (2018)
9. Heller, S., Gsteiger, V., Bailer, W., 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. IJMIR **11**(1), 1–18 (2022)
10. Jegou, H., et al.: Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 117–128 (2010)
11. Li, Y., Song, Y., Cao, L., Tetreault, J., et al.: TGIF: a new dataset and benchmark on animated gif description. In: Proceedings of IEEE CVPR 2016 (2016)
12. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
13. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 185–201. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_12
14. Markatopoulou, F., Moumtzidou, A., Galanopoulos, D., et al.: ITI-CERTH participation in TRECVID 2017. In: Proceedings of TRECVID 2017 Workshop, USA (2017)
15. Pittaras, N., Markatopoulou, F., Mezaris, V., Patras, I.: Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In: Amsaleg, L., Guðmundsson, G.Þ, Gurrin, C., Jónsson, B.Þ, Satoh, S. (eds.) MMM 2017. LNCS, vol. 10132, pp. 102–114. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51811-4_9

16. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML) (2021)
17. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019)
18. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C – a research video collection. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., Vrochidis, S. (eds.) MMM 2019. LNCS, vol. 11295, pp. 349–360. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05710-7_29
19. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., et al.: CrowdHuman: a benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)
20. Tan, M., Le, Q.V.: Efficientnet: rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
21. Wang, X., et al.: Vatex: a large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of IEEE/CVF ICCV 2019, pp. 4581–4591 (2019)
22. Xu, J., Mei, T., et al.: MSR-VTT: a large video description dataset for bridging video and language. In: Proceedings of IEEE CVPR 2016, pp. 5288–5296 (2016)