Proc. MMM 2022, Springer LNCS vol. 13142. Author's accepted version. The final authenticated version is avalable online at https://doi.org/10.1007/978-3-030-98355-0_50

VERGE in VBS 2022

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

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

Abstract. This paper presents VERGE, an interactive video search engine that integrates multiple retrieval methodologies and also combines them with reranking and fusion techniques. Moreover, a user interface, implemented as a Web application, enables users to formulate queries, view the top retrieved shots and watch the respective videos, before submitting a shot to a VBS task, all in an efficient and easy manner.

1 Introduction

VERGE is an interactive video search engine that provides a multitude of retrieval capabilities along with a friendly and simple user interface (UI) for submitting queries and viewing the top results. With a long-standing participation in the Video Browser Showdown (VBS) competition [17], VERGE is constantly reformed in order to better tackle the competition's "Ad-Hoc Video Search" (AVS) and "Known Item Search - Visual/Textual" (KIS-V, KIS-T) tasks. This year, existing search modalities are improved, e.g. visual similarity and face detection, others are extended, e.g. to more concepts and activities, while novel concept-based late fusion techniques are introduced. Moreover, the VERGE UI has been adapted accordingly to support all the different search approaches.

The paper is structured as follows: Section 2 presents the overall framework of the VERGE engine, Section 3 continues with the detailed description of the retrieval components, Section 4 illustrates the UI and some usage scenarios, and Section 5 concludes with the future work.

2 The VERGE Framework

As depicted in Fig. 1, the VERGE framework consists of three layers. The first layer contains the various retrieval modalities that are applied on the competition's datasets, i.e. V3C1 and V3C2 [16], and in most cases their results are prestored in a database. The second layer refers to the services that can accept queries and respond with the top results, i.e. the most relevant shots/videos. The third layer is the Web Application UI that allows users to formulate their queries, calls the services and visualises the retrieved results.

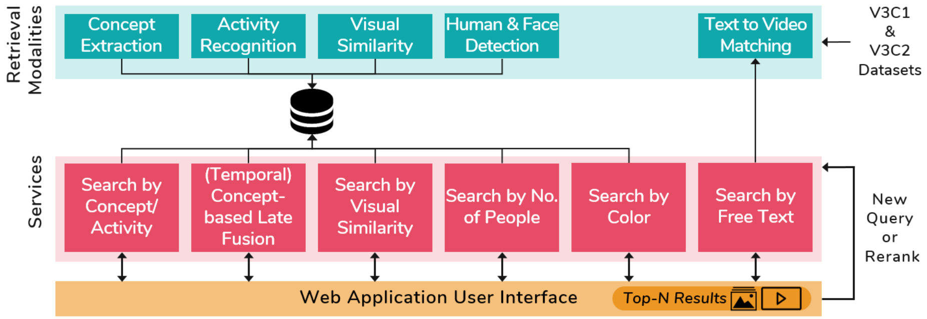


Fig. 1. The VERGE Framework

3 Retrieval Modalities

3.1 Concept-Based Retrieval

This module annotates each keyframe with a pool of concepts, which comprises 1000 ImageNet concepts, a selection of 300 concepts of the TRECVID SIN task [14], 500 event-related concepts, 365 scene classification concepts, 580 object labels, 30 style-related concepts 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 three pre-trained models that employ different DCNN architectures, i.e. the EfficientNet-B3, EfficientNet-B5 [19] and InceptionResNetV2. To obtain scores for the subset from the TRECVID SIN task, we trained and employed two models based on the EfficientNet-B1 and EfficientNet-B3 architectures on the official SIN task dataset. For the event-related concepts we used the pre-trained model of EventNet [9]. Regarding the extraction of the scene-related concepts, we utilized the publicly available VGG16 model, fine-tuned on the Places365 dataset [23]. Object detection scores were extracted using models pre-trained on the established MS COCO and Open Images V4 datasets, with 80 and 500 detectable objects, respectively. For the style-related concepts we employed the pre-trained models of [20]. 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 various text similarity measures between all concepts' labels. After manual inspection of the text analysis results we formed groups of very similar concepts for which we create a common label and assign the max score of its members.

3.2 Spatio-temporal Activity Recognition

This module aims to extract human-related activities for each shot to enable filtering functionalities using the labels of the activities. A sorted list of 700 predefined human-related activities and the corresponding prediction scores are

Proc. MMM 2022, Springer LNCS vol. 13142. Author's accepted version. The final authenticated version is avalable online at https://doi.org/10.1007/978-3-030-98355-0_50

extracted using a DCNN architecture. Specifically, a 3D-CNN, similarly to [10], which efficiently learns spatio-temporal activity representations of human-related activities using a 3D-ResNet architecture. The model's pre-trained weights are learned using Kinetics-700 dataset [3], while its architecture comprises 152 layers. The spatio-temporal dimensions of the model's input correspond to 112×112 spatial size with 16 frames temporal duration.

3.3 Visual Similarity Search

This module retrieves visually similar content using DCNNs and two approaches will be tested. The first one uses a 1024 vector from a fine-tuned GoogleNet architecture [15] and an IVFADC index database vector that is created with these vectors [11]. The second one involves the use of a new Bayesian-based image retrieval method [13]. As input, we use the visual features extracted from the fc-7 layer of VGG-16 for each shot. Then, we compute the semantic affinities of them, transform them into a probability distribution and project them in the Hamming space via minimizing the Kullback-Leibler divergence. Next, Bayesian regression is used for learning the projection of visual features to the learnt hash codes. For achieving fast retrieval, we use k-NN on GPU [8] between hash codes.

3.4 Human and Face Detection

This module aims to detect and count humans and human faces in each frame shot, so that the user can easily distinguish the results of single-human or multi-human activities. The detection can be performed in crowd-center scenes, where partial occlusions among humans or between humans and objects are possible to occur. To deal with this, a dataset named CrowdHuman [18] is selected, which includes annotations for the humans and their corresponding faces. Using this dataset, an optimal, in terms of speed and accuracy, object detector YoloV4 [1] is fine-tuned, intended to extract the detected human silhouettes and faces, making it possible to count the total number of them.

3.5 Text to Video Matching Module

The text to video matching module inputs a complex free-text query along with a set of video shots and returns a ranked list with the most relative video shots w.r.t. to the input textual query. For this, we utilize the attention-based dual encoding network presented in [6]. This network is trained to transform video-caption samples into a new joint embedding space. In this embedding space, a straightforward comparison between free-text queries and video or image instances is feasible. The network consists of two similar sub-networks [4] in parallel, one for the video shot and one for the natural language sentence. Each sub-network consists of multiple levels of encoding, i.e. using mean-pooling, attention-based [6], bi-GRU sequential model, and CNN layers. Following the state-of-the-art approach [4–6], the improved marginal ranking loss [5] is used to train the entire network. Following [7], we effectively combine the results delivered from multiple trained models. Regarding the training data, we use a combination of four large-scale video caption datasets: MSR-VTTT [22], TGIF [12], ActivityNet [2] and Vatex [21]. The ResNet-152 deep network trained on the ImageNet-11k dataset is used as initial video shot representation.

3.6 Concept-based Late Fusion

This module combines two or more visual concepts (Section 3.1) and generates a sorted list of shots using a late fusion approach. First, separate lists of shot probabilities for each concept are created. Then, the intersection at shot level of the concepts is computed, which is reranked by an objective function that respects the principle that the higher the concept probabilities or the more relevant the shots are, the higher their scores are. To achieve this, we compute the difference of the concept probabilities for all possible concept pairs and then we apply the inverse exponential function to them.

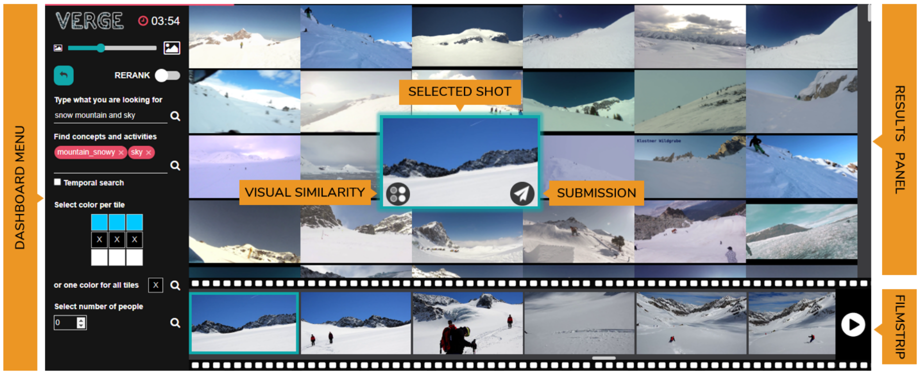
3.7 Temporal Late Fusion

This module incorporates two or more visual concepts (Section 3.1) and generates a sorted list of unique videos using a late fusion approach. At first, for each concept a separate list of shot probabilities is created. Then, the intersection of concepts per video is computed and only the first valid ordered tuple of each video is kept. Those shots are reranked through an objective function, which preserves the concept-based late fusion principles (Section 3.6) and also favours the ordering of the concepts by using a weighting function.

4 User Interface and Interaction Modes

The VERGE UI (Fig. 2) is provided as a Web application that allows its users to easily make queries, view the results, watch the corresponding videos and submit their selection during the VBS contest. The main characteristics that we aim for the UI to have are compactness (minimum usage of space for maximum functionality), intuitiveness (when a user sees it, they know exactly what to do), and efficiency (the speed of using it is not frustrating).

The VERGE interface contains three basic parts: the dashboard menu on the left, the results panel in the center, and the filmstrip on the bottom. The *dashboard menu* includes a countdown timer, indicating the remaining time for submission during an active VBS task, an undo button to return to the previous results, a switch button to select between a new query and reranking, and, finally, the search modules. The first search option is a text input field, where the user is able to describe with free text query what they are looking for (Section 3.5). The second option offers a long list of concepts and activities (Sections 3.1, 3.2) that the user can search with scrolling or autocomplete. Multiple selection is supported for late fusion (Section 3.6) and a checkbox can activate temporal



Type what you are looking for
snow mountain and sky

Find concepts and activities
mountain_snowy × sky ×

☐ Temporal search

Select number of people
0

Fig. 2. The VERGE User Interface

fusion (Section 3.7). The third option is search by color, where the user is able to paint certain tiles of the image to be sought in a 3x3 grid. The fourth option is to retrieve shots with a set number of people appearing in them (Section 3.4). The top retrieved shots of all search modalities are visualised in the *results panel* in a grid view. Hovering over a shot shows two additional buttons: one for applying the fifth and final retrieval method, i.e. visual similarity (Section 3.3), and another one for submitting a shot to a contest task. When a shot is clicked, the *filmstrip* is updated with all the frames of the video it belongs to, while the the button on the right can play the video in a popup.

To illustrate the usability of VERGE in the VBS contest, we discuss briefly here some cases that tackle actual queries of VBS 2021. For an AVS task that requests shots of "two women walking and talking", the user can fuse the concept "female_person" with the activities "walking" and "talking", and then rerank by setting the desired number of people to two. For a KIS-V task that shows a snow mountain and the sky, the user can search by color, painting the top tiles light blue and the bottom tiles white, and when a relevant image appears, they can apply visual similarity in order to get more results and find the specific shot (Fig. 2). Finally, for a KIS-T task that reads "a hand opening the window of a mountain hut", the user can use this exact sentence for free-text search and the expected shot can be found within the top twenty results.

5 Future work

While a constant goal is to improve the developed retrieval algorithms both in terms of performance and time efficiency, the experience of VBS 2022 will drive the next modifications in the available search options as well as the VERGE UI.

Acknowledgements This work has been supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-825079 MindSpaces, H2020-780656 ReTV, and H2020-832921 MIRROR.

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: Proc. of IEEE CVPR 2015. pp. 961–970 (2015)
3. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
4. Dong, J., Li, X., Xu, C., Ji, S., He, Y., et al.: Dual encoding for zero-example video retrieval. In: Proc. of IEEE CVPR 2019. pp. 9346–9355 (2019)
5. Faghri, F., Fleet, D.J., et al.: VSE++: Improving visual-semantic embeddings with hard negatives. In: Proc. of BMVC 2018 (2018)
6. Galanopoulos, D., Mezaris, V.: Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In: Proc. of ACM ICMR 2020 (2020)
7. Galanopoulos, D., Mezaris, V.: Hard-negatives or Non-negatives? A hard-negative selection strategy for cross-modal retrieval using the improved marginal ranking loss. In: Proc. of IEEE/CVF ICCVW 2021 (2021)
8. Garcia, V., Debreuve, E., Barlaud, M.: Fast k Nearest Neigh-bor Search using GPU. In: Proc. of ACM ICMR 2008. ACM (2008)
9. Guangnan, Y., Yitong, L., Hongliang, X., et al.: Eventnet: A large scale structured concept library for complex event detection in video. In: Proc. of ACM MM 2015 (2015)
10. Hara, K., et al.: Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In: Proc. of IEEE CVPR 2018 (2018)
11. Jegou, H., et al.: Product quantization for nearest neighbor search. IEEE transactions on pattern analysis and machine intelligence **33**(1), 117–128 (2010)
12. Li, Y., Song, Y., Cao, L., Tetreault, J., et al.: TGIF: A new dataset and benchmark on animated gif description. In: Proc. of IEEE CVPR 2016 (2016)
13. Lin, Z., Ding, G., Hu, M., Wang, J.: Semantics-Preserving Hashing for Cross-View Retrieval. In: Proc. of IEEE CVPR 2015 (2015)
14. Markatopoulou, F., Moumtzidou, A., Galanopoulos, D., et al.: ITI-CERTH participation in TRECVID 2017. In: Proc. of TRECVID 2017 Workshop. USA (2017)
15. Pittaras, N., et al.: Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In: MMM. pp. 102–114. Springer (2017)
16. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C–a research video collection. In: Proc. of MMM 2019. pp. 349–360. Springer (2019)
17. Schoeffmann, K., Lokoč, J., Bailer, W.: 10 years of video browser showdown. In: Proc. of ACM MM 2021. pp. 1–3 (2021)
18. 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)
19. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
20. Tan, W.R., et al.: Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In: 2016 IEEE ICIP. pp. 3703–3707. IEEE (2016)
21. Wang, X., et al.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proc. 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: Proc. of IEEE CVPR 2016. pp. 5288–5296 (2016)
23. Zhou, B., Lapedriza, A., et al.: Places: A 10 million image database for scene recognition. IEEE transactions on PAMI **40**(6), 1452–1464 (2017)