VERGE in VBS 2024

Nick Pantelidis[1](✉), Maria Pegia[1,2], Damianos Galanopoulos[1], Konstantinos Apostolidis[1], Klearchos Stavrothanasopoulos[1], Anastasia Moumtzidou[1], Konstantinos Gkountakos[1], Ilias Gialampoukidis[1], Stefanos Vrochidis[1], Vasileios Mezaris[1], Ioannis Kompatsiaris[1], and Björn Þór Jónsson[2]

[1] Information Technologies Institute/Centre for Research and Technology Hellas, Thessaloniki, Greece
{pantelidisnikos,mpegia,dgalanop,kapost,klearchos_stav,moumtzid, gountakos,heliasgj,stefanos,bmezaris,ikom}@iti.gr
[2] Reykjavik University, Reykjavik, Iceland
{mariap22,bjorn}@ru.is

Abstract. This paper presents VERGE, an interactive video content retrieval system designed for browsing a collection of images extracted from videos and conducting targeted content searches. VERGE incorporates a variety of retrieval methods, fusion techniques, and reranking capabilities. It also offers a user-friendly web application for query formulation and displaying top results.

1 Introduction

VERGE is an interactive system for video content retrieval that offers the ability to perform queries through a web application in a set of images extracted from videos, display the top ranked results in order to find the appropriate one and submit it. The system has continuously participated in the Video Browser Showdown (VBS) competition [17] since its inception. This year, we enhanced the performance of existing modalities for better performance and introduced new ones based on the lessons learned from the VBS 2023 [17]. Moreover, we made some subtle yet effective enhancements to the web application, all aimed at providing a better user experience.

The structure of the paper is as follows: Sect. 2 presents the framework of the system, Sect. 3 describes the various retrieval modalities, Sect. 4 presents the UI and its features, concluding with Sect. 5 that outlines the future work.

2 The VERGE Framework

Figure 1 visualizes the VERGE framework that is composed of four layers. The first layer consists of the various retrieval modalities that most of them are applied beforehand to the datasets and their results are stored in a database. The second layer contains the Text to Video Matching service that uses the

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14557, pp. 356–363, 2024.
https://doi.org/10.1007/978-3-031-53302-0_32

corresponding modality to return results. The third layer consists of the API endpoints that read data either from the database or the services and return the top results for each case. The last layer is the Web Application that sends queries to the API and displays the results.

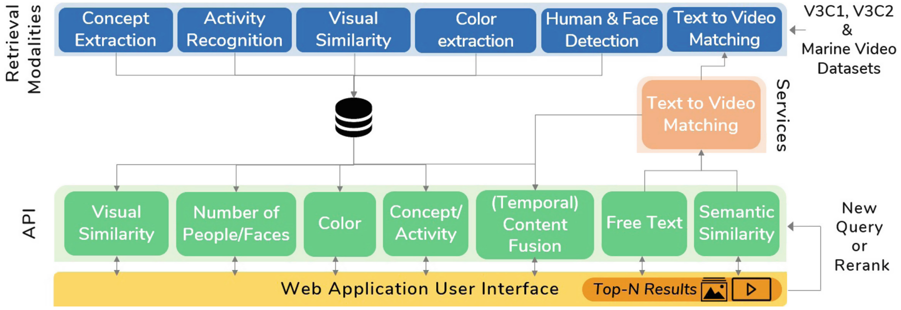


Fig. 1. The VERGE Framework

VERGE integrates the V3C1+V3C2 dataset [23], the marine video dataset [26] and the new dataset for this year, the LapGyn100 dataset (surgeries in laparoscopic gynecology). The shot segmentation information was provided for the initial two datasets. For the third dataset, OpenCV[1] was used to segment videos into coherent shots, with each shot lasting about 10 s. A sampling ratio of 5 was applied to reduce frame retention.

The main updates realized this year include: the addition of semantic similarity search and the late fusion using the text to video matching module which also was improved in terms of speed, the use of advanced architecture for the concept extraction module, and the application of preprocessing steps related to human/face detection for activity recognition.

3 Retrieval Modalities

3.1 Concept-Based Retrieval

This module annotates each shot's keyframe with labels from a pool of concepts consisting of 1000 ImageNet concepts, a subset of 298 concepts from the TRECVID SIN task [19], 365 scene classification concepts, 580 object labels, and 22 sports classification labels. To generate annotation scores for the ImageNet concepts, we adopted an ensemble approach, averaging the concept scores from two pre-trained models, the ViT_B_16 [13] and EfficientNet_V2_L [25]. For the subset of concepts from the TRECVID SIN task, we trained and employed a model based on the EfficientNet-B4 architecture using the official SIN task dataset. In the case of scene-related concepts, we adopted the same approach,

[1] https://opencv.org/.

averaging the concept scores from WideResNet [29] and ResNet50 [10] models, pre-trained on the Places365 [30] dataset. Object detection scores were extracted using models pre-trained on the MS COCO [16] and Open Images V4 [14] datasets, which include 80 and 500 detectable objects, respectively. To label sports in video frames, a custom dataset was created using web images related to sports and used to train a model based on the EfficientNetB3 architecture. For a more concise listing representation of the concept-based annotations, we applied the sentence-BERT [22] text encoding framework to measure the text similarity among all concept labels. After analyzing the results, we manually grouped very similar concepts, assigning them a common label and taking the maximum score among their members.

3.2 Spatio-Temporal Activity Recognition

This module processes each video shot to extract one or more human-related activities. Thus, enhances the filtering capabilities by adding activity label options to the list of concepts. A list of 400 pre-defined human-related activities was extracted for each shot using a 3D-CNN architecture deploying a pipeline similar to [8] that modifies the 3D-ResNet-152 [9] architecture. During inference, the shots were fed to the model to fit the model's input dimension, $16 \times 112 \times 112 \times 3$, and the extracted activities were sorted according to their scores, from higher to lower. Finally, a post-processing step was applied to remove false-positive activity predictions by excluding those that weren't human-related, as the human and face detection module (Sect. 3.6) did not predict human bodies or faces.

3.3 Visual Similarity Search

This module uses DCNNs to find similar content based on characteristics extracted from each shot. These characteristics are derived from the final pooling layer of the GoogleNet architecture [21], serving as a global image representation. To enable swift and effective indexing, we create an IVFADC index database vector with these feature vectors, following the technique proposed in [12].

3.4 Text to Video Matching Module

This module inputs a complex free-text query and retrieves the most relevant video shots from a large set of available video shots. We extend the $\mathrm{T} \times \mathrm{V}$ network [6] into the $\mathrm{T} \times \mathrm{V} + \mathit{Objects}$. The $\mathrm{T} \times \mathrm{V}$ network consists of two key subnetworks, one for the textual and one for the visual stream and utilizes multiple textual and visual features along with multiple textual encoders to build multiple cross-modal common latent spaces. The $\mathrm{T} \times \mathrm{V} + \mathit{Objects}$ network utilizes visual information at more than one level of granularity. Specifically, it extends $\mathrm{T} \times \mathrm{V}$ by introducing an extra object-based video encoder using transformers.

Regarding the training, a combination of four large-scale video caption datasets is used (e.g. MSR-VTTT [28], TGIF [15], ActivityNet [2] and Vatex

[27]), and the improved marginal ranking loss [4] is used to train the entire network. We utilize four different textual features: i) Bag-of-Words (bow), ii) Word2Vec model [20], iii) Bert [3] and iv) the open_clip ViT-G14 [11] model. These features are used as input to the textual sub-network ATT presented in [5]. As a second textual encoder, we consider a network that feedforwards the open_clip features. Similarly to the textual sub-network, we utilize three visual encoders to extract frame-level features: i) R_152, a ResNet-152 [10] network trained on the ImageNet-11k dataset, ii) Rx_101 a ResNeXt-101 network, pre-trained by weakly supervised learning on web images followed and fine-tuned on ImageNet [18] and iii) the open_clip ViT-G14 [11] model. Moreover, we utilize the object detector and the feature extractor modules of the ViGAT network [7] to obtain the initial object-based visual representations.

3.5 Semantic Similarity Search

The module is used to find video shots that are semantically similar, meaning they may not be visually similar but carry similar contexts. To find semantically similar shots, we utilize and adapt the $\mathrm{T} \times \mathrm{V} + \mathit{Objects}$ of Sect. 3.4 used in the text-to-video matching module. The $\mathrm{T} \times \mathrm{V} + \mathit{Objects}$ model consists of two key sub-networks (textual and visual). It associates text with video by creating new common latent spaces. We modify this model to convert a query video and the available pool of video shots into the new latent spaces where a semantic relationship is exposed. Based on these new video shot representations we retrieve the most related video shots with respect to the video shot query.

3.6 Human and Face Detection

This module aims to detect and report the number of humans and human faces in each keyframe of each shot to enhance user filtering by providing capabilities for distinguishing the results of single-human or multi-human activities. The detections of both human silhouettes (bodies) and human faces (heads) were extracted using YoloV4 [1]. MS COCO [16] datasets' weights were used and fine-tuned using the CrowdHuman dataset [24], where partial occlusions among humans or between humans and objects can occur in crowd-centre scenes.

3.7 Late Fusion Approaches for Conceptual Text Matching

This section explores two late fusion approaches for combining conceptual text and shot detection in video analysis. Thus, concepts from Sect. 3.1 or sentences from Sect. 3.4 serve as text information. These methods create a list of shots, utilizing sequential or intersection information from texts, and sorting them using

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

where $P = \{p_i\}_{i=1}^{i=n}$ are the shot probabilities computed by modules from Sect. 3.1 or Sect. 3.4.

Conceptual Text Late Fusion. This method focuses on assembling a list of shots that encompass specific sentences or concepts from the textual content. It identifies shots that appear in both text queries, indicating shared context and content. The process involves developing independent lists of shot probabilities for each text and subsequently calculating their intersection at shot level. The final list is then sorted using Eq. 1.

Temporal Text Late Fusion. The method returns a list of tuples of shots, with each element corresponding to the same video and containing all queried texts in a particular order. This approach incorporates text queries and sorts them using the same late fusion method (Eq. 1). The process includes creating a list of shot probabilities for each text, calculating their intersection at video layer, retaining the first tuple of each video while respecting the order of texts, and finally sorting the shots using the Eq. 1.

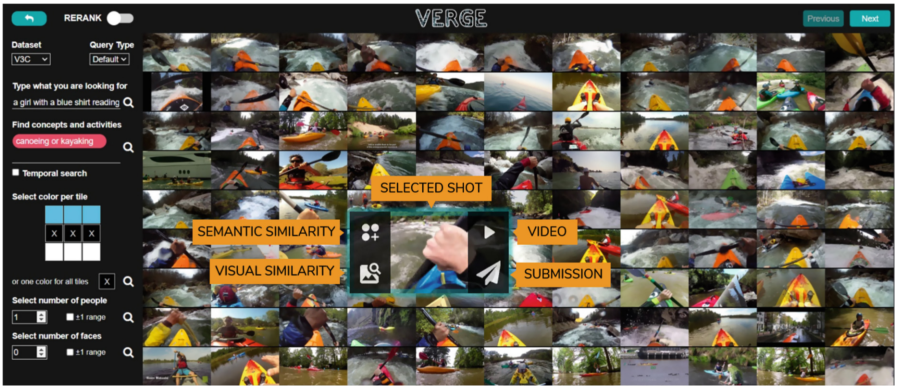
3.8 Semantic Segmentation in Videos

The module is designed to provide precise pixel-level annotations of surgical tools and anatomical structures within each shot frame. This capability enables the user to distinguish objects of interest against the surgical laparoscopic background easily and to locate the relevant scenes. To achieve this, we employ the state-of-the-art DCNN architecture, LWANet. The model's initial weights are acquired through pre-training on the ImageNet dataset, followed by fine-tuning on the domain-specific EndoVis2017 dataset tailored for laparoscopic surgery scenarios. During inference, the model processes every frame of a shot, generating a binary mask that precisely outlines the areas of segmented objects.

4 User Interface and Interaction Modes

The VERGE UI (Fig. 2) is a web-based application that allows users to create and execute queries, utilizing the retrieval modalities that were described above. The top results from the queries are displayed so as the users can browse through them, watch the corresponding videos and submit the appropriate data.

The UI is composed of three main parts. The header on the top, the menu on the left and the results panel that takes all the remaining space. The header contains, from left to right, the back button for restoring previous results, the rerank button for returning the intersection of the current results and the results of the new performed query, the logo on the center and the pagination buttons on the right for navigating through the pages. The first options on the menu are the dataset selector, which allows you to choose the dataset for conducting queries, and the query type where when the AVS option is selected, yields a single result per video. The initial search module is the free text search, allowing users to input anything in the form of free text. Subsequently, the concept/activity-based



Dataset: V3C
Query Type: Default
Type what you are looking for: a girl with a blue shirt reading
Find concepts and activities: canoeing or kayaking
□ Temporal search
Select color per tile
or one color for all tiles
Select number of people: 1 □ ±1 range
Select number of faces: 0 □ ±1 range

Fig. 2. The VERGE User Interface

search offers pre-extracted values for users to choose from. Multiple selection is also supported for late fusion as well as temporal fusion if the corresponding checkbox is checked. The color search is achieved by selecting the color for a part of the image in a 3 × 3 grid or by selecting a color for the whole image. Finally, there is possibility to search based on the count of people or faces discernible in the image, alongside the choice to conduct a flexible search by opting to include results with either an additional or one less person/face. Moreover, when a user hovers over a result/image, four buttons appear (Fig. 2). From the top-left corner and in a clockwise order, these are: the button that returns images with similar context, the one that returns visually similar images, the video play button and the button for sending this shot to the competition's server.

To showcase the VERGE functionality, we present three use cases. In the first case, we are looking for shots of canoeing and kayaking so we can select from the concepts/activities the activity with the name canoeing or kayaking. In the second use case, we want to find a girl with a blue shirt that reads a book outdoors. We can use this exact description in the free text search. The last use case involves locating a skier within a scene that encompasses both the sky and the snow. Therefore, we can color the relevant regions of the image accordingly (Fig. 2), and then reorganize the outcomes based on the visible individuals.

5 Future Work

This year, the newly added LapGyn100 dataset introduced unique challenges due to its nature. Our upcoming months of research will focus on exploring methods tailored to address these challenges that involve further study of images processing techniques and changes to the UI. Finally, the changes realised are expected to enhance the UI's user-friendliness for both expert and novice users while maintaining its functionality.

Acknowledgements. This work has been supported by the EU's Horizon 2020 research and innovation programme under grant agreements 101070250 XRECO, 883302 ISOLA, 883484 PathoCERT, 101021866 CRiTERIA, and the Horizon research and innovation programme under grant agreement 101096405 UP2030.

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., Escorcia, V., Ghanem, B., Carlos Niebles, J.: ActivityNet: a large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–970 (2015)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
4. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: Proceedings of the British Machine Vision Conference (BMVC) (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 the ACM International Conference on Multimedia Retrieval (ICMR 2020). ACM (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, pp. 627–643. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25069-9_40
7. Gkalelis, N., Daskalakis, D., Mezaris, V.: ViGAT: bottom-up event recognition and explanation in video using factorized graph attention network. IEEE Access **10**, 108797–108816 (2022). https://doi.org/10.1109/ACCESS.2022.3213652
8. Gkountakos, K., Touska, D., Ioannidis, K., Tsikrika, T., Vrochidis, S., Kompatsiaris, I.: Spatio-temporal activity detection and recognition in untrimmed surveillance videos. In: Proceedings of the 2021 International Conference on Multimedia Retrieval, pp. 451–455 (2021)
9. Hara, K., et al.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of IEEE CVPR 2018 (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Ilharco, G., et al.: OpenCLIP (2021). https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below
12. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 117–128 (2010)
13. Kolesnikov, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale (2021)
14. Kuznetsova, A., et al.: The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vis. **128**(7), 1956–1981 (2020)

15. Li, Y., et al.: TGIF: a new dataset and benchmark on animated GIF description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4641–4650 (2016)
16. 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
17. Lokoč, J., et al.: Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th VBS. Multimed. Syst. **29**, 3481–3504 (2023)
18. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 181–196 (2018)
19. Markatopoulou, F., Moumtzidou, A., Galanopoulos, D., et al.: ITI-CERTH participation in TRECVID 2017. In: Proceedings of TRECVID 2017 Workshop, USA (2017)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, Workshop Track Proceedings, ICLR 2013 (2013)
21. 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
22. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv preprint arXiv:1908.10084 (2019)
23. 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
24. 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)
25. Tan, M., Le, Q.: Efficientnetv2: smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106. PMLR (2021)
26. Truong, Q.T., et al.: Marine video kit: a new marine video dataset for content-based analysis and retrieval. In: Dang-Nguyen, D.T., et al. (eds.) MMM 2023. LNCS, vol. 13833, pp. 539–550. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-27077-2_42
27. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: VATEX: a large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4581–4591 (2019)
28. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: a large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5288–5296 (2016)
29. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
30. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)