# A VR Interface for Browsing Visual Spaces at VBS2021

Ly-Duyen Tran[1], Manh-Duy Nguyen[1], Thao-Nhu Nguyen[1], Graham Healy[1], Annalina Caputo[1], Binh T. Nguyen[2,3], and Cathal Gurrin[1]

[1] Dubin City University, Ireland
[2] AISIA Research Lab
[3] Vietnam National University, Ho Chi Minh University of Science

**Abstract.** The Video Browser Showdown (VBS) is an annual competition in which each participant prepares an interactive video retrieval system and partakes in a live comparative evaluation at the annual MMM Conference. In this paper, we introduce Eolas, which is a prototype video/image retrieval system incorporating a novel virtual reality (VR) interface. For VBS'21, Eolas represented each keyframe of the collection by an embedded feature in a latent vector space, into which a query would also be projected to facilitate retrieval within a VR environment. A user could then explore the space and perform one of a number of filter operations to traverse the space and locate the correct result.

**Keywords:** Video Browser Showdown · Interactive Retrieval · Virtual Reality · Eolas

## 1 Introduction

As the volume of multimedia data increases, there is a need for valid experimental comparisons between competing approaches. The Video Browser Showdown (VBS) meets this need by providing a means of comparing interactive search systems using known-item search tasks over large video collections (partly) in front of a live audience [8]. In this work, we introduce an experimental VBS prototype developed by a DCU and HCMUS-based team participating for the first time. The experimental system called Eolas is an interactive retrieval system that provides a novel Virtual Reality (VR) interface to large multimedia libraries. The system comprises two main components, the back-end embedding and clustering techniques for data storage/retrieval and the front-end voice and gesture-controlled VR interface for interaction.

This paper's contribution is in describing Eolas, where users can explore the dataset and seek a specific video in a user-friendly virtual environment without the visual interface constraints imposed by using a desktop screen. Moreover, our voice control protocol allows users to avoid the laborious task of typing text when in VR. The embedding scheme, combined with a clustering step, also helps users get a better experience analyzing the dataset.

## 2 Related Systems

The VBS has witnessed many video retrieval systems that participated in various approaches during the previous nine years. This section briefly describes several top-performing teams in the competition last year to indicate the current state-of-the-art.

The main goal of SOM-Hunter [5], the top-performing system in 2020, was to combine intuitive text search and browsing with more advanced optional options for experts. The top-ranked frames for a query were displayed using a self-organizing map (SOM) to visualize the high-dimensional data. Additionally, users could choose relevant frames for relevance feedback and they could repeat this process until a suitable result was found, or they could start-over at any time with a new query. The runner-up and the third-placed systems were "VIRET" [6] and the "vitrivr" system [9], respectively. Both tools supported novel query generation methodologies, such as Query-by-Sketch, which facilitated users to draw a query and Query-by-Example, allowing a user to search for visually similar keyframes to a selected one. They still had the typical approach of using a textual query in which each image frame was annotated by its semantic concepts such as detected objects or scene text appearing within the frame. VIRET aimed their interface at novice users by supporting easier query re-formulation, while 'vitrivr" focused on their storage database structure to reduce searching time. A fourth system to note was Exquisitor [4], which was a notable and novel prototype based on large-scale interactive learning. It relies on chosen positive and negative examples of visualized keyframes to learn a simple relevance model. The retrieval approach is supported with an efficient index.

VR systems for interactive retrieval have not yet received significant research attention. However, one notable approach is the VR platform for multimodal Lifelog data [2], which was best placed in the 2018 Lifelog Search Challenge [3] by focusing on providing novice users with an immersive and easy to search and browse large archives. Eolas will build upon the user-friendly nature of the VR-lifelog tool [2] by supporting easy querying and implementing a state-of-the-art latent space index support targeted browsing of the collection, as shown to be effective by previous participants. The 360-degree view afforded by a VR platform can help users by providing a larger display area.

## 3 An Overview of Eolas

Eolas can be viewed in terms of the two prominent components: indexing the dataset in which we annotated all frames into a latent feature space and our unique user interaction support for the VR environment. The overview of Eolas can be depicted in Figure 1.

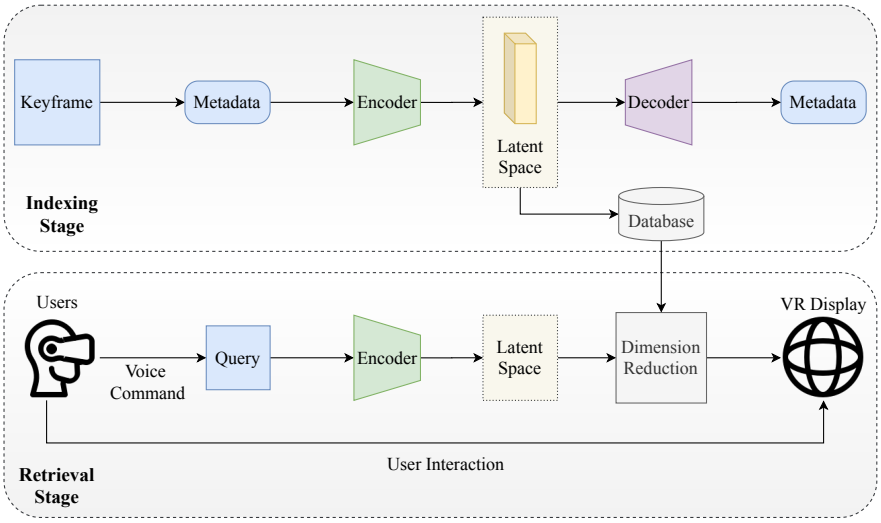


**Fig. 1.** Eolas Workflow. In the indexing stage, each frame was embedded into a latent vector space based on its metadata by training an Autoencoder network and then was stored in the index. Users wearing a VR device could input a semantic query by using our voice command protocol. The query was also converted into the same latent space as the database. All feature vectors were then transformed into 3D space for visualisation in the VR environment to find the result.

### 3.1 Source Data

In this task, we use the first part of the Vimeo Creative Commons Collection dataset[1] (V3C) called V3C1 [1], a wide-ranging video collection. V3C1 dataset includes 7,475 videos (1,000 hours), categorized into different content categories ranging from Food, Fashion, Art to Instructional videos. The videos were segmented into shots which were represented by over a million keyframes. Additionally, we integrated various forms of provided metadata, such as shot captions, visual concepts, and text extracted from the shots as an input for the retrieval engine. We also amended additional metadata to support user filters, such as a face detector's output that enumerated the number of faces in the shot, a color histogram, audio detection, and camera motion vectors. The sources are presented in Table 1.

### 3.2 Search Engine

Since each keyframe was represented by the various forms of metadata presented above, we approached the video retrieval problem as a semantic challenge and every keyframe was annotated by combining the three primary sources of textual information (captions, concepts, and text).

[1] https://sigmm.hosting.acm.org/2019/07/06/the-v3c1-dataset-advancing-the-state-of-the-art-in-video-retrieval/

**Table 1.** List of all metadata

| Name | Definition |
|---|---|
| caption | Text generated from Google Cloud Speech-to-Text API |
| concept | Detection of objects in keyframes |
| text | Detection of text in keyframes (OCR) |
| color | Detection of dominant color in keyframes |
| faces | Detection of faces in keyframes |
| histogram | Histogram data on bitrate, resolution, duration and upload dates |
| camera motion | Detection of static or moving cameras |
| speech detector | Detection of music or silence in videos |

**Index Construction** The semantic data of a frame was first embedded into a feature space using the Glove model [7], followed by a concatenation step to be fused into a single vector. We then built an Autoencoder network in which an encoder module was used to create a lower-dimensional space to store representative features in the vector and the decoder component needed to ensure these encoded attributes could be reverted to its original value. It is noted that the Autoencoder network would be integrated with the RNN sequence model to cope with the issue of arbitrary length textual descriptions. After the training stage, a keyframe could be annotated by its meaningful features and stored in our index for fast retrieval. An extra supervised learning clustering technique was applied to discover similar groups of keyframes, which could be useful while users solve a task. This step was done once while indexing data and the groups were then stored in our index.

**Retrieval** A textual query was also fed into the encoding module to be converted into the same vector space with keyframe embedded features in the retrieval stage. We then applied a dimensional reduction algorithm to visualize the data in the 3D virtual space, where the users' initial position was based on the query's projection into space. The video keyframes were also illustrated in this space through their 3D converted vector features. The ocular distances within this space indicated the similarity between the query and frames. Keyframes were clustered into groups from which several images would be shown. When in the virtual environment, users could select relevant keyframes. All images from their groups with the query's features were then transformed into a 1D line, which we called similarity path. This path could help users to focus on these keyframes rather than the entire dataset.

### 3.3 User Interaction

There are two phases within which the user interacts with the system. For each task, the user uses voice commands prompting the system to start a new search from the spoken query. Using ASR, the spoken inquiry is converted to text query (which is embedded in the latent space) and is used to bring the VR user to a

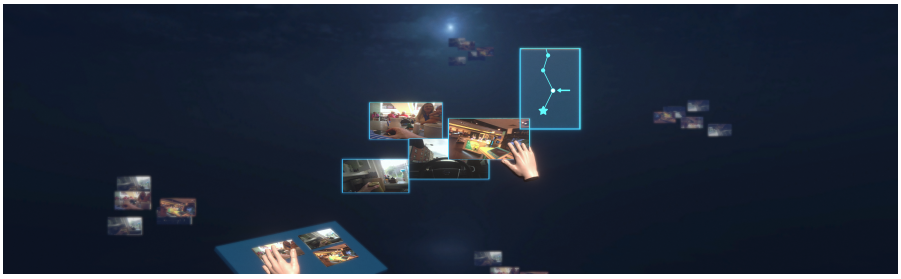

**Fig. 2.** A VR Visual Interface prototype

starting cluster in the 3D space. In the second phase of interaction, the user can navigate the space to locate the required videos. The user has the option to select the most suitable videos and use a voice command to ask the system to take him/her to the closest video cluster to the selected videos. The videos are ranked and shown to the user in decreasing rank order on a similarity path. At any time, the user can (1) restart the search, (2) go back to the previous cluster, (3) go to any point in the similarity path on the minimap, or (4) choose more videos in order to rerank the similarity path.

**VR Visual Interface** Our proposed visual interface is designed to be accessible for novice users. The user interface is kept minimal, consisting of two parts: a minimap showing the current similarity path, which is always floating on the right of the user's head, and a saved section on the left side of the user. All videos in the 3D space are represented by tangible objects that the user can grab and move in the space or put in the saved section. The user can travel around the environment, either using voice commands to navigate the similarity path on the minimap or through a traditional VR teleportation system by choosing the nearby clusters visible from the current location.

## 4 Conclusions

This paper introduces Eolas, which is an experimental video retrieval system built on the VR platform. By utilising the wide and surrounding view in the virtual environment, this tool can easily visualize the grouped videos in the 3D dimensional space based on their encoded features and provide users with an intuitive and novel access mechanism for large video archives.

**Acknowledgments** This publication has emanated from research supported in party by research grants from Science Foundation Ireland under grant numbers 18/CRT/6223 and 18/CRT/6224.

## References


1. 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. p. 334–338. ICMR '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3323873.3325051, https://doi.org/10.1145/3323873.3325051
2. Duane, A., Gurrin, C., Huerst, W.: Virtual reality lifelog explorer: lifelog search challenge at acm icmr 2018. In: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge. pp. 20–23 (2018)
3. Gurrin, C., Schoeffmann, K., Joho, H., Leibetseder, A., Zhou, L., Duane, A., Nguyen, D., Tien, D., Riegler, M., Piras, L., et al.: Comparing approaches to interactive lifelog search at the lifelog search challenge (lsc2018). ITE Transactions on Media Technology and Applications **7**(2), 46–59 (2019)
4. Jónsson, B.Þ., Khan, O.S., Koelma, D.C., Rudinac, S., Worring, M., Zahálka, J.: Exquisitor at the video browser showdown 2020. In: International Conference on Multimedia Modeling. pp. 796–802. Springer (2020)
5. Kratochvíl, M., Veselý, P., Mejzlík, F., Lokoč, J.: Som-hunter: Video browsing with relevance-to-som feedback loop. In: Ro, Y.M., Cheng, W.H., Kim, J., Chu, W.T., Cui, P., Choi, J.W., Hu, M.C., De Neve, W. (eds.) MultiMedia Modeling. pp. 790–795. Springer International Publishing, Cham (2020)
6. Lokoc, J., Kovalcík, G., Soucek, T.: VIRET at video browser showdown 2020. In: Ro, Y.M., Cheng, W., Kim, J., Chu, W., Cui, P., Choi, J., Hu, M., Neve, W.D. (eds.) MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 11962, pp. 784–789. Springer (2020). https://doi.org/10.1007/978-3-030-37734-2_70
7. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
8. Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., Vrochidis, S.: Interactive video retrieval in the age of deep learning - detailed evaluation of vbs 2019. IEEE Transactions on Multimedia pp. 1–1 (2020)
9. Sauter, L., Amiri Parian, M., Gasser, R., Heller, S., Rossetto, L., Schuldt, H.: Combining boolean and multimedia retrieval in vitrivr for large-scale video search. In: Ro, Y.M., Cheng, W.H., Kim, J., Chu, W.T., Cui, P., Choi, J.W., Hu, M.C., De Neve, W. (eds.) MultiMedia Modeling. pp. 760–765. Springer International Publishing, Cham (2020)