Videofall - A Hierarchical Search Engine for VBS2022

Thao-Nhu Nguyen[1], Bunyarit Puangthamawathanakun[1], Graham Healy[1], Binh T. Nguyen[2,3], Cathal Gurrin[1], and Annalina Caputo[1]

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

Abstract. In this paper, we introduce a multi-user hierarchical video search tool called VIDEOFALL. Our objective, in the Video Browser Showdown (VBS) 2022, is to explore if VIDEOFALL interactive video retrieval under time constraints is a useful approach to take, given the overhead of requiring multiple users. It is our conjecture that combining different skills of normal users can support a master user to retrieve target videos efficiently. The system is designed on top of the CLIP pre-trained model and the video keyframes are embedded into a vector space in which queries would also be encoded to facilitate retrieval.

Keywords: Video Browser Showdown · Interactive Video Retrieval · Hierarchical Engine · Multi-user Search Engine

1 Introduction

The motivation for VIDEOFALL is that each individual has distinct abilities and skills that can uniquely inform the retrieval process for the video retrieval challenge. Hence, our conjecture is that combining the efforts of multiple users into one collaborative searching tool could be an interesting approach for the VBS.

In this paper, we propose VIDEOFALL, a hierarchical interactive video search engine that enables multiple users to interact simultaneously, all with the aim of finding potentially relevant pieces of video content from a large archive. We utilise a pre-trained Contrastive Language–Image Pre-training (CLIP) model [4] as the backbone of the retrieval approach, which supports a group of users to search using various techniques, such as colour, OCR or textual descriptions. Afterward, upon finding a potentially relevant item, those users pass it onto a master user for final review before selecting it as submittable or not. The master user will consequently handle the higher-level task of validating all non-overlapping results from the database and sending back the final result to the server.

2 Related Research

The Video Browser Showdown (VBS), an annual competition in the Multimedia Modeling Conference (MMM) since 2012, is known for supporting a live experi-

mental comparison between video retrieval systems. All competitors are required to resolve two main types of tasks: Known-Item search (KIS) and Ad-hoc Video Search (AVS). AVS simulates the scenario of finding as many correct scenes as possible from the given description. Meanwhile, KIS is concerned with locating just one specific scene fitting with the provided description. There are two subtasks of KIS including (1) visual KIS, where the clip containing the target scene will be displayed and (2) textual KIS with an increasingly detailed query description being shown.

There have been a number of notable relevant approaches taken during the most recent VBS challenges. The third release of VISIONE [1] allowed user to query by combining different types of representations including text, objects, colour, visual examples, and their spatial relationships. For each specific type of information, different AI-based techniques were separately employed in order to extract as many features as possible. SOM-Hunter [9] has proven to be a high-performing system in which the scenes were ranked and visualized on high-dimensional data by using a self-organising map (SOM). Aside from employing the traditional bag-of-words model for describing objects, SOM-Hunter also enabled seeking targets by multiple text queries to capture the positional relationships between objects; a user had the ability to specifically describe objects or events in a sub-region of the whole scene. While most of the teams have designed traditional 2D retrieval systems, vitrivr-VR [3] introduced an interactive Virtual-Reality-based engine that co-embedded text and video into the same high-dimensional space. Additional information from Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) were also attached to that 3D feature space. Users could use both textual and speech-to-text queries in order to retrieve the target frame. Then, the result was presented in a 2D scrolling list according to the ranked list of scenes. Different from vitrivr-VR, the main goal of EOLAS [8] - another VR system first introduced last year - was to explore the latent feature space where all frames were embedded. Based on the cosine distance between encoded vectors, the group of videos would be visualized on the wide and surrounding 3D view in the virtual environment. As a result, user was able to directly interact with the system.

There are a few multiple-user system that have developed before. Schoeffmann at el. [7] provided a collaborative search system in which each individual user received a specific view from the web interface. The authors used Feature Map, a similarity-based map of keyframes, as their main interface. Furthermore, all users shared the same synchronized view of inspection actions in a cooperative heatmap. Then, the collaborative retrieved results would be re-ranked to finalise the submittable result. Unlinke Schoeffmann's system [7], ours primarily utilise the distinct point of view to explore different aspects of one event. Therefore, we provide all normal users the same UI. The final result is only submitted by the master user after evaluating all possible frame received from the unique normal users.

3 An Overview of VIDEOFALL

In general, the search engine of VIDEOFALL consists of two fundamental components: (1) the indexer, where all keyframes extracted from the given videos are encoded to latent space using the CLIP image encoder; and (2) the retrieval engine, in which the input query is also embedded and then ranked based on the cosine distance. Meanwhile, the User Interface (UI) is designed to ease multiple-user access. Figure 1 indicates the VIDEOFALL workflow overview, while the User Interaction flow and the UI protocols of the system are illustrated in Figure 2 and Figure 3 respectively.

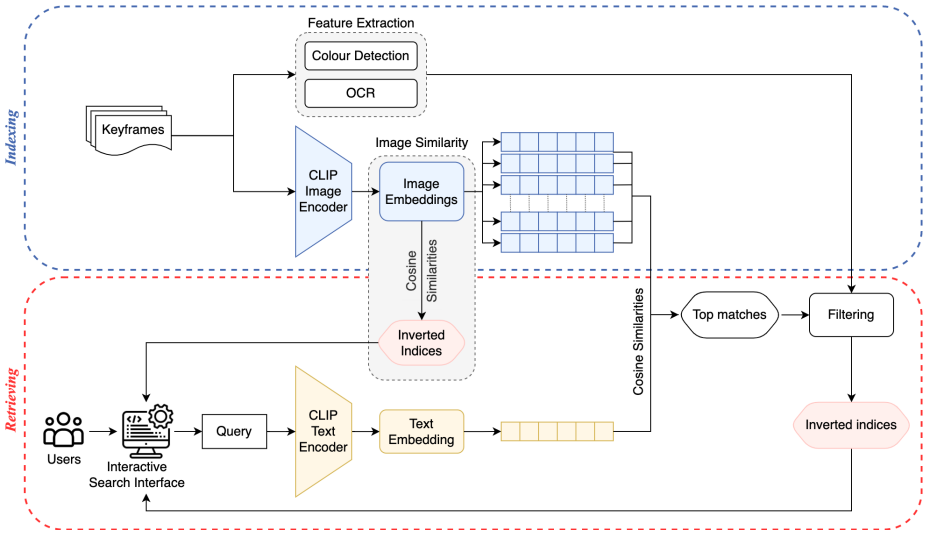

Fig. 1: VIDEOFALL Workflow

3.1 Source Data

This year's competition will combine two parts of the Vimeo Creative Commons Collection dataset (V3C) [6], which is comprised of V3C1 [2] and V3C2 [5]. These two extensive collections of videos, with different topic categories ranging from Arts, Fashion, Technology to Food, were already segmented into shots represented by pre-extracted keyframes. We did not alter the provided shot boundaries or keyframes. The dataset details are presented on Table 1.

3.2 System Overview

As shown in Figure 1, the overall system consists of two main stages: indexing and retrieving. At indexing time, colour and OCR are extracted from the original keyframes in order to support retrieving in the next stage.

Table 1: Dataset details

Name	No. videos	No. keyframes	No. hours of content	Size (TB)
V3C1	7475	1,082,659	1,000	1.3
V3C2	9760	1,425,454	1,300	1.6
Total	**17,235**	**2,508,113**	**2,300**	**2.9**

Feature Extraction Our goal is to determine the globally dominant colour of one frame which would be used as a filter to narrow the search space. That colour would belong to one of the 32 basic colour bins on the palette. Thinking of each pixel as a data point in 3D space, we apply the KMeans clustering algorithm to the set of points. The colour has the nearest distance to the group's center point value would be assigned as the label of that group. The denser the group is, the more that colour dominates the frame. After that, both the top 5 densest groups in ascending order and the extracted OCR will be combined as metadata to support queries.

Indexing We will leverage the pre-trained CLIP model [4] from OpenAI corporation, which aims to make connections between the visual and textual representations. CLIP text and image encoders will be trained on the full-sentence description of the image instead of just one label, resulting in a deeper understanding of the dataset. Furthermore, training on a couple of large-scale datasets (e.g. ImageNet, Youtube-BB, ObjectNet, ...) supports the model learning various aspects of images. In VIDEOFALL, the CLIP image encoder will be exploited to convert all keyframes into feature vectors in a latent space.

Retrieval Once users input a textual query, it will be fed into the pre-trained text encoder to be encoded into a single vector having the same dimension as the image embedding above. The pairwise cosine distances between all image text pairs are computed before being ranked for later use. Then, we will get the ranked list of the top-most fitting images to the description. The same process is applied to the image embedding vectors to create an inverted index of similar images. These indices enable the direct retrieval of relevant images given a sample scene, also known as image similarity.

3.3 User Interface

We have introduced two simple UIs for two main phases in user interaction. The UI for the first phase, illustrated in Figure 3a, is designed to be easily accessible for the searches performed by normal users. For each task, users translate the provided topic description into the query for the UI. After obtaining the list of matching images in reverse order of relevance to the query, this is refined to remove overlapping images among users. At this stage, the users have 2 options,

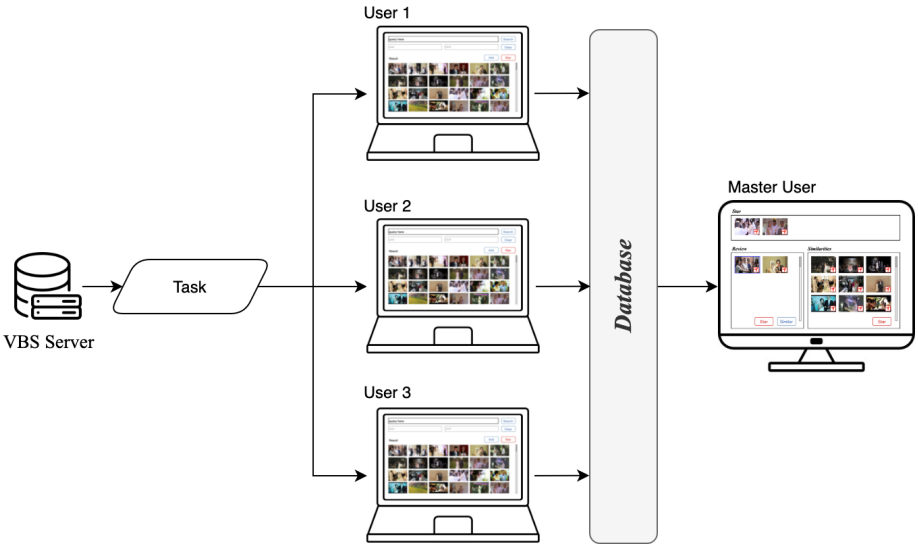

Fig. 2: The User Interaction Flow

either add an image to the Review list for validation by the master user in the next stage, or put an image in the Star list if they find the correct shot. In the second phase, the UI (Figure 3b) is developed specifically for the master user who will take responsibility for reviewing candidates and generating the final submission. From the Review list, the master user can use visual similarity to find similar images to any image in the Review list, or they can just directly sumbit from the Star list. It is noteworthy that the master user can submit the final results at any time.

(a) (b)

Fig. 3: The two User Interface protocols: (a) UI for the normal users and (b) UI for the master user.

4 Conclusion

In this paper, we have introduced a multi-user interactive system, VIDEOFALL, built on top of the CLIP model. We decided to make a multi-user system because different people can bring different skillsets to the retrieval process. VIDEOFALL aims at capturing the wide range of users' understandings of the queries and the aggregation of their searching may help the team to find required content faster and more effectively.

Acknowledgments This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 18/CRT/6223, and 13/RC/2106_P2 at the ADAPT SFI Research Centre at DCU. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

References

1. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: VISIONE at Video Browser Showdown 2021. In: Lokoč, J., Skopal, T., Schoeffmann, K., Mezaris, V., Li, X., Vrochidis, S., Patras, I. (eds.) MultiMedia Modeling. pp. 473–478. Springer International Publishing, Cham (2021)
2. 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
3. Heller, S., Gasser, R., Illi, C., Pasquinelli, M., Sauter, L., Spiess, F., Schuldt, H.: Towards Explainable Interactive Multi-modal Video Retrieval with Vitrivr. In: Lokoč, J., Skopal, T., Schoeffmann, K., Mezaris, V., Li, X., Vrochidis, S., Patras, I. (eds.) MultiMedia Modeling. pp. 435–440. Springer International Publishing, Cham (2021)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. CoRR **abs/2103.00020** (2021), https://arxiv.org/abs/2103.00020
5. Rossetto, L., Schoeffmann, K., Bernstein, A.: Insights on the V3C2 dataset. CoRR **abs/2105.01475** (2021), https://arxiv.org/abs/2105.01475
6. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C - a research video collection. CoRR **abs/1810.04401** (2018), http://arxiv.org/abs/1810.04401
7. Schoeffmann, K., Primus, M.J., Muenzer, B., Petscharnig, S., Karisch, C., Xu, Q., Huerst, W.: Collaborative feature maps for interactive video search. In: Amsaleg, L., Guðmundsson, G., Gurrin, C., Jónsson, B., Satoh, S. (eds.) MultiMedia Modeling. pp. 457–462. Springer International Publishing, Cham (2017)
8. Tran, L.D., Nguyen, M.D., Nguyen, T.N., Healy, G., Caputo, A., Nguyen, B.T., Gurrin, C.: A VR Interface for Browsing Visual Spaces at VBS2021. In: Lokoč, J., Skopal, T., Schoeffmann, K., Mezaris, V., Li, X., Vrochidis, S., Patras, I. (eds.) MultiMedia Modeling. pp. 490–495. Springer International Publishing, Cham (2021)

9. Veselý, P., Mejzlík, F., Lokoč, J.: SOMHunter V2 at Video Browser Showdown 2021. In: Lokoč, J., Skopal, T., Schoeffmann, K., Mezaris, V., Li, X., Vrochidis, S., Patras, I. (eds.) MultiMedia Modeling. pp. 461–466. Springer International Publishing, Cham (2021)