Vibro: Video Browsing with Semantic and Visual Image Embeddings

Konstantin Schall, Nico Hezel, Klaus Jung, and Kai Uwe Barthel

HTW Berlin, University of Applied Sciences - Visual Computing Group
Wilhelminenhofstraße 75, 12459 Berlin, Germany
konstantin.schall@htw-berlin.de
http://visual-computing.com/

Abstract. *Vibro* represents a powerful tool for interactive video retrieval and browsing and is the winner of the Video Browser Showdown 2022. Following the saying of "never change a winning system" we did not change any of the underlying concepts nor added any new features. Instead, we focused on improving the three existing cornerstones of the software, which are text-to-image search, image-to-image search and browsing results with 2D sorted maps. The changes to these three parts are summarized in this paper, and in addition, an overview of the *AVS*-mode of *vibro* is given.

Keywords: Content-based Video Retrieval · Exploration · Visualization · Image Browsing · Visual and Textual co-embeddings

1 Introduction

The Video Browser Showdown (VBS) [4] is a international competition where participating teams have to solve three task categories: *visual Known-Item Search* (v-KIS), *textual Known-Item Search* (t-KIS) and *Ad-Hoc Video Search* (AVS). In the upcoming 2023 event, the data will consist of the combination of three datasets: V3C1 (7475 video files with a duration of 1000 hours), V3C2 (9760 videos with 1300 hours), and a relatively small dataset of diving videos, which are very similar and highly redundant [11].

The large variety of tasks and data lead to very specific requirements for the participating systems. First, the sheer amount of data, with nearly 3 TB of videos, requires a very memory efficient solution and implementation of the software. Second, the systems have to be able to formulate queries in different modalities to solve each of the task categories and browse the results. Third, the underlying analysis of video data must be sufficiently generalizable to work with the diverse videos of the V3C datasets and the low variety of the scuba diving videos.

In the last version of our *vibro* system, we mainly focused on improving the user interface and introduced a CLIP model [8] for joint-embeddings to support full-text search. However, since CLIP embeddings encapsulate information of

images more on a semantic than a detailed visual level, we used a second embedding for image-based queries. Thanks to these improvements, *vibro* took first place in the overall competition.

This paper presents the next iteration of *vibro* and is structured as follows. For completeness, we want to showcase the main features of *vibro* in the first section. Following the saying of "never change a winning system", we have left the main principles and functionalities of the previous version untouched and only updated the underlying embeddings for textual and image-based queries which is highlighted in the following section. In the last part, we focus on the *AVS* mode of the system, since it uses its own user interface and ranking model and has not been previously described in great detail.

2 GUI and Features

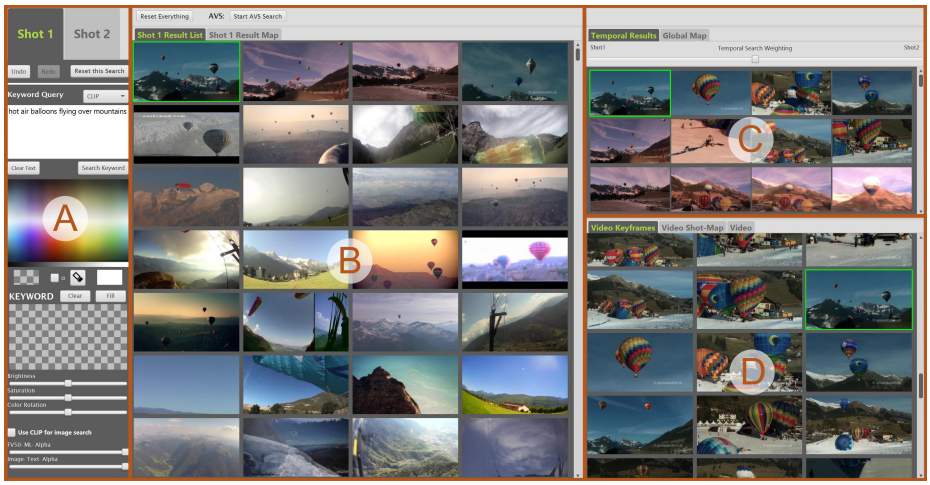

Fig. 1: Graphical user interface for the current version of *Vibro*

Figure 1 gives an overview of *vibro's* user interface. Part A is reserved for query formulation and the results are presented in part B. Here, the top 4000 frames are displayed either in a 2D-sorted map or in a simple descending multi-column list, ordered by relevance to the current query. Searches can be formulated for two consecutive shots by selecting the tab-buttons at the top of part A. If both tabs contain a query, the queries are used in a temporal search. Those results are sorted in part C as a list of five consecutive shots, each representing a segment of a video. Part D is reserved for single video views. One tab displays each shot of the currently selected video, another tab contains a video player. The last view (Global Map) is a further tab in Section C and allows exploring the whole video collection with an image similarity graph [5].

The system supports queries of three modalities: sketch, text and image-by-example. These modalities can be mixed with adjustable weights. Sketch-based queries can be formulated either by drawing on the blank canvas in part A or by selecting an image from any of the other parts and subsequent modification with the drawing tools. Rich-Text queries are supported by employing a CLIP trained joint-embedding model.

Each text input is encoded into a vector by the textual encoder of the model and then a simple similarity search is performed over all shot-vectors, generating a score for each shot. The embeddings for all shots are extracted once and stored locally. Image queries can be initiated by double clicking any shot presented in one of the views. We use a separate set of embeddings produced by a deep neural network, that has been fine-tuned for the specific case of content-based image retrieval in this step and rank all video-shots by its relevance to the query by an exhaustive similarity search. Additionally, out-of-database queries are possible by dragging an image file or URL on the drawing canvas.

For temporal search, we try to arrange sequences of keyframes in such a way that one keyframe matches the query of the first search tab and one of the three following keyframes of the same video matches the query of the second tab. To do this, the similarity values of the search results of tab 1 are combined with the highest of each of the three search results of tab 2 to calculate the harmonic mean of both scores to describe the temporal search result score.

3 Improvements of the Current Version

The three main improvements can be summarized as follows. First, we updated the CLIP model from a ResNet [3] based visual encoder to a visual transformer [2] version (*ViT-L@336*). Second, the content-based retrieval model was significantly improved and last, we updated the sorting algorithm to FLAS [1], which not only slightly improved the arrangement quality, but also led to faster sorting speed.

When we started working on the 2022 version of *vibro*, the largest and best performing version of CLIP, namely *ViT-L@336*, was not publicly available. For this reason we fell back to the second best model, which was the *ResNet50x16* version. Additionally, we conducted heuristic experiments on compression with PCA and concluded that everything below 512 dimensions would harm the text-to-image retrieval qualities. These findings also align with [7]. Since even a modest compression affects the order of results, we have omitted it completely this year. However, the embeddings are still quantized to bytes which is the only post-processing step. With these changes, text-to-image retrieval performs better compared to the last version of *vibro*, since we are not only using a better visual encoder in the first place, but also omit compression, which slightly reduced the retrieval quality.

One of the most challenging aspects of the *VBS* is the high diversity of the video data and because of that, all video-shots have to be encoded by well generalizing deep neural networks. CLIP was trained with over 400 million of image

Table 1: Comparison of Different Encoders and their Embeddings

Encoder	GPR1200 mAP	Data Type	Dim	Image Size
CLIP Vibro22	71.2	Byte	512	384
CLIP Vibro23	76.1	Byte	768	336
CBIR Vibro22	73.4	Bit	1024	384
CBIR Vibro23	86.1	Bit	1024	336
CBIR Vibro23	87.2	Float	1024	336

and text pairs and as shown by Radfort et al. [8], generalizes to a lot of different tasks. However, since the objective of CLIP is to pair text with images, it is not specialized for the task of content-based image retrieval (CBIR). For example, a nearest neighbor search with a picture of a burger on a plate in a restaurant will yield images of popular fast-food restaurant logos and buildings. Semantically, both the burger and the restaurant incorporate the concept of fast-food and therefore are similar but the particular object instances are different. Networks that specifically have been trained for the task of CBIR, might perform better for image-to-image queries. In the last version of *vibro* a Swin-L network [6] was used for CBIR and we exchanged that model for a ViT-L network, pretrained with CLIP and finetuned with a combination of publicly available image datasets such as ImageNet21k [9] and Google Landmarks v2 [12]. The final training set had over 22 million images from 168 thousand categories. We then evaluated all models with the *mean average precision* score of the GPR1200 dataset [10], which was designed as a benchmark for general-purpose CBIR solutions. Table 1 compares the CLIP and CBIR networks and shows that the current version is significantly superior in image-to-image retrieval settings. The final visual embeddings have 1024 dimensions and are binarized with a threshold of 0, which only marginally harms the retrieval performance but allows us to use the *Hamming distance* for similarity search and reduces the memory requirement by factor 32 compared to the floating point version.

Two of the featured UI parts benefit from the third change we made this year. The first one is the 2D-map arrangement view of the 4000 most relative shots to the query found in section B and the second one is the "Global Map" navigation tab in section C. In both cases video-shots are arranged on a 2D-grid with FLAS [1]. FLAS produces arrangements approximately ten times faster than a traditional SOM and 10% faster than a SSM, while providing better sorting quality.

4 AVS Mode

Whereas the goal of the *v-KIS* and t-KIS tasks is to find one specific video, as many as possible shots fitting a specific textual description have to be submitted for the *ad-hoc video search* (*AVS*) part of the competition. Since videos can only be submitted from part D of the main user interface, we decided to implement

a separate interface for the ad-hoc video search. Figure 1 and 2 show how an example task of "find shots with hot air balloons flying over mountains" could be solved with *vibro*. First, the main UI is used to execute a textual query. Next, one of the presented video-shots is selected and the "Start AVS Search" button on the very top of part B of Figure 1 opens the *AVS* window shown in Figure 2. Presented are 42 video-shots selected by similarity search with the CBIR embeddings and the selected shot from the main UI as a query. Additionally, all shots are filtered to have an adjustable temporal distance to any of the already presented shots to omit to many results from the same video. The user now has to mark all relevant shots with a click on the image and submits them by pressing the "Send" button (left of Figure 2). For the next iteration all marked video-shots are used to perform a multi-image similarity search. The scores are merged with a minimum function over each of the query images and the found shots are again filtered temporally, taking into account all previously submitted video-shots. The result of the second iteration is shown in Figure 2 on the right. This procedure can be repeated until no more relevant images are presented in the result list. The user can then close the *AVS* window, select another initial video-shot from the main UI and open a new *AVS* window. Again, all previously submitted and presented shots are taken into account, unless everything has been reset with the corresponding button.

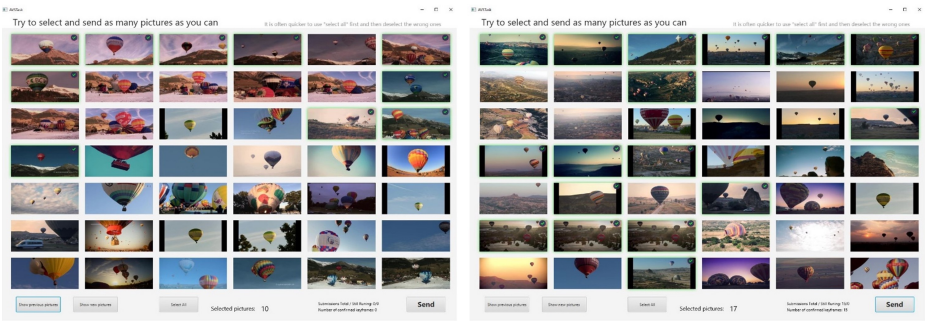

Fig. 2: AVS mode of the current version of *Vibro*

5 Conclusion

This paper presents the 2023 version of *vibro*, a powerful video browsing tool that will participate in the upcoming VBS (Video Browser Showdown). The main improvements can be summarized as enhancements in text-to-image and image-to-image retrieval and also faster and better arrangements of the search results on 2D-grids for easier browsing.

References

1. Barthel, K.U., Hezel, N., Jung, K., Schall, K.: Improved evaluation and generation of grid layouts using distance preservation quality and linear assignment sorting (2022). https://doi.org/10.48550/ARXIV.2205.04255
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR (2020)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
4. Heller, S., Gsteiger, V., Bailer, W., Gurrin, C., Jónsson, B., Lokoc, J., Leibetseder, A., Mejzlík, F., Peska, L., Rossetto, L., Schall, K., Schoeffmann, K., Schuldt, H., Spiess, F., Tran, L., Vadicamo, L., Veselý, P., Vrochidis, S., Wu, J.: Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown. Int. J. Multim. Inf. Retr. **11**(1), 1–18 (2022). https://doi.org/10.1007/s13735-021-00225-2, https://doi.org/10.1007/s13735-021-00225-2
5. Hezel, N., Barthel, K.U.: Dynamic construction and manipulation of hierarchical quartic image graphs. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. p. 513–516. ICMR '18, Association for Computing Machinery, New York, NY, USA (2018)
6. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)
7. Lokoč, J., Souček, T.: How many neighbours for known-item search? In: Reyes, N., Connor, R., Kriege, N., Kazempour, D., Bartolini, I., Schubert, E., Chen, J.J. (eds.) Similarity Search and Applications. pp. 54–65. Springer International Publishing, Cham (2021)
8. 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)
9. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) (2015)
10. Schall, K., Barthel, K.U., Hezel, N., Jung, K.: Gpr1200: A benchmark for general-purpose content-based image retrieval. In: MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I. p. 205–216. Springer-Verlag, Berlin, Heidelberg (2022)
11. Truong, Q.T., Vu, T.A., Ha, T.S., Lokoč, J., Tim, Y.H.W., Joneja, A., Yeung, S.K.: Marine video kit: A new marine video dataset for content-based analysis and retrieval. In: MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023. Springer (2023)
12. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval. In: CVPR (2020)