AVSeeker: An Active Video Retrieval Engine at VBS2022

Tu-Khiem Le[1(✉)], Van-Tu Ninh[1], Mai-Khiem Tran[2,3,4], Graham Healy[1], Cathal Gurrin[1], and Minh-Triet Tran[2,3,4]

[1] Dublin City University, Dublin, Ireland
tukhiem.le4@mail.dcu.ie, tu.ninhvan@adaptcentre.ie
[2] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[3] John von Neumann Institute, VNU-HCM, Ho Chi Minh City, Vietnam
[4] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. Exploring video clips in a vast collection of videos is a difficult task. It is necessary to provide an efficient system for users to express the information need for sought events in that video collection. Thus, we propose to develop AVSeeker – an active video retrieval engine – to assist users in finding appropriate moments in videos with two main query types: textual descriptions and visual examples. The main characteristic of AVSeekeris that we change the retrieval engine from a passive system to an active one, which narrows the search space by gaining clues from users through an interactive relevance feedback manner. The AVSeeker is based on the LifeSeeker system from the annual Lifelog Search Challenge with the addition of an interactive relevance feedback via concept recommendation, enriched temporal concepts, and query-by-sketch functionality.

Keywords: Video retrieval system · Interactive relevance feedback retrieval system · Query by sketch

1 Introduction

Retrieving items from a large multimedia collection (known-item search) is a non-trivial problem and many research benchmarking challenges have been established to progress the domain. The Video Browser Showdown (VBS) [13] is one such challenge for video retrieval, which is held annually to evaluate different interactive video search engines. It is a live competition that provides a fair and accurate comparison of many interactive video retrieval system. In the VBS, a query can be in the form of a text in which new information is extended gradually to mimic the human's information seeking process, or in the form of a short video so that users can formulate the query themselves.

In this work, we present the Active Video Seeker (AVSeeker) - an interactive video search engine - with the core infrastructure inherited from our interactive lifelog retrieval system (LifeSeeker [8,9]), which competed in the Lifelog Search

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13142, pp. 537–542, 2022.
https://doi.org/10.1007/978-3-030-98355-0_51

Challenge (LSC) [5] for several years. AVSeeker is further enhanced with new relevant feedback search mechanism using concept recommendation and visual similarity search using sketch.

The main contributions of this paper are: (1) AVSeekeris introduced which incorporates a novel search mechanism in which the system actively suggests a list of relevant concepts related to the queries to refine the retrieval results rather than having the system passively return paginated candidate results in response to the user input; (2) A new temporal concept enrichment profess is proposed which enriches the concepts extracted from a video clip by integrating action/activity recognition; (3) A new query-by-sketch mechanism is described, which allows a user to retrieve the images by sketching the scene, which can help to bridge the limitation in the number of available visual concepts from the concept detectors.

2 Related Work

The most recent edition of the Video Browser Showdown (VBS) was in 2021 and we will now introduce some of the notable systems that participated in 2021. Vitrivr [6], achieved the highest score in the challenge. The authors of the system inherited some video pre-processing techniques of its predecessor [12] to enrich the metadata of the video for efficient retrieval, including Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), concept enrichment using Faster-RCNN model pre-trained on the Openimages V4 dataset. Additionally, the system was also enhanced with the temporal querying mechanism and an efficient index structure for similarity search, including text and image search, which were claimed to increase the overall score of the searched results.

The video search engine from the SOMHunter team [7] also provided the same mechanisms of both temporal query and image search, however, with a different implementation. However, their main contribution was a customised relevance feedback mechanism to be used with self-organising maps (SOMs) [7] where users provided positive and negative examples for better localization of the exact result. The keywords of example frames were also used as a basis for keyword refinement after the relevance feedback process. This proposed feature showed its efficiency in VBS'20 by achieving the highest score in the competition.

For AVSeeker, we also integrated some typical features of such state-of-the-art interactive video retrieval systems, and propose a new interactive video retrieval technique via user feedback through concept suggestions to narrow the searching corpus. The description of our system is discussed detailed in Sect. 3.

3 AVSeeker – An Active Video Retrieval System

AVSeeker is developed based on our current lifelog retrieval system, which is LifeSeeker 2.0 [9], by directly extending the system to support video search. The main differences between the two retrieval system are discussed in Sects. 3.2, 3.3, and 3.4.

3.1 LifeSeeker – An Overview of Our Predecessor Retrieval System

LifeSeeker is an interactive retrieval system designed for the LSC challenge that emphasises on still-image search based on a given text query. The LSC dataset contains a collection of lifelog images in the first person perspective and the challenge requires the search engine to find one image which matches the query as quickly and accurately as possible. LifeSeeker operates on a free-text-search mechanism in which the system matches a given text query with a set of lifelog images. This was achieved by converting the text query and images into the same vector space and compare the similarity between the query's vector and the images' vectors. The images in the dataset were processed by multiple detectors to extract visual concepts, including objects [1], scene categories and scene attributes [18], and scene texts [2], for matching with the text query. An Elasticsearch [4] server was deployed to aid the process of indexing and retrieving the relevant images. Additionally, a Bag-of-Words matching mechanism was also implemented and served as a extra option for searching. Lastly, LifeSeeker also provided a visual similarity search function so that the user can retrieve the images which share the similar scene and structure.

Apart from inheriting the LifeSeeker's core functionality, we introduce our first video search prototype AVSeeker that is customised for the video search task. Given the differences between the LSC and VBS datasets (video keyframes and not lifelog images), AVSeeker is further enhanced with new search mechanisms described in detail in the following sections.

3.2 Active Video Search via Relevance Feedback Mechanism Using Concept Recommendation

We inherit the idea of the Akinator[1] application, which is a game trying to guess of which characters or objects the player is thinking, using a series of questions generated by the machine and obtaining corresponding multiple choice answers from the user. Based on this guessing game, we try to transform our passive search system into an active retrieval engine that supports the user during the searching process. In conventional passive search systems, the user needs to think of relevant concepts related to the information needed based on the description of the query, which depends heavily on the ability of the user and ultimately relies too much on the user to know which concepts are likely to assist in finding relevant content. In contrast, for an active retrieval engine, the mutual interaction between the user and the retrieval engine is more important. In AVSeeker, the system acts as an assistant for the user during the searching progress by suggesting related concepts to the query based on an initial input given by the user to narrow the set of relevant items.

In detail, the process of active video search via relevance feedback mechanism using concept recommendation follows three steps as follows:

1. At first, the user inputs some initial concepts that are best described the scene into the text search box, which is the conventional approach.

[1] https://en.akinator.com.

2. The system follows a concept pre-processing pipeline to retain important concepts (e.g. objects, animals, environment) and inputs them into ConceptNet[2] [14] to enhance the general context-understanding of the concepts (e.g. location, type, usage, related terms) for the construction of concept suggestion. An initial ranked list of relevant items which are related to corresponding concepts is returned to the user.
3. A list of scene-relevant concept is recommended to the user to gather the user feedback. Thereby, the result set is then quickly narrowed to a smaller set of relevant scenes.

3.3 AVSeeker's Extended Features from LifeSeeker

In LifeSeeker, we already have the mechanism to retrieve events based on concept matching. The more accurate concepts that can be extracted from the visual data, the better the performance. With each video clip, in addition to the information in the still image, we can also exploit information about the motion between the image frames. For instance, a video of a car running on a busy street and the other one with a car running on a highway can be distinguished via the movement trajectory of the vehicle. Additionally, other videos related to human understanding can also be recognized via the detection of the actions of the characters in the video. In this way, we find more concepts from a video clip for our search engine to work on. To achieve this, we apply the solution of finding objects based on their visual features and especially movement trajectories to the algorithm proposed by Nguyen et al. [11]. Furthermore, we employ the solutions from Vo et al. in [15,16] for finding action proposals, then we classify those candidates into known activities. This is similar to the process of detecting objects from images: we obtain a region proposal for an object, then we classify the region into known objects from training data.

3.4 Visual Similarity Search by Sketch

In the VBS challenge, the queries are not only expressed in text, but they also can be given as a visual example. For these queries, we integrate conventional visual similarity search approaches including Bag-of-Features model using Scaled-Invariant-Feature-Transform (SIFT) [10] and Speeded-up Robust Features (SURF) [3], region-based color clustering, Nearest-Neighbor search of the latent space representation of video frames. Additionally, we also integrate a Query-by-Sketch function into the AVSeeker by extending previous work in [17]. Once the user finishes a sketch and submits to the AVSeeker server, a pre-trained model extended from [17] is employed to adapt the sketch features to the video frame domain for similarity mapping. This process returns a ranked-list of videos in descending order of Euclidian distance in the latent space of the sketch feature.

[2] https://conceptnet.io/.

4 Conclusion

In this paper, we introduce our initial work of an active video retrieval – the AVSeeker – competing in VBS 2022. The AVSeeker inherits most of the features from its predecessor lifelog retrieval engine – LifeSeeker, and enhances the user interaction for large-scaled visual data search by proposing a concept-recommendation system which plays a role of an interactive relevance feedback to improve the precision of the search result. The AVSeeker also supports Query-by-Sketch function for queries that require visual similarity search.

Acknowledgement. This work was supported by the ADAPT Centre (Grant 13/RC/2106; 13/RC/2106_P2) and Insight Centre for Data Analytics (Grant SFI/12/RC/2289_P2) funded by Science Foundation Ireland Research Centres Programme and co-funded by the European Regional Development Fund.

References

1. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
2. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9365–9374 (2019)
3. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**, 346–359 (2008)
4. Gormley, C., Tong, Z.: Elasticsearch: The Definitive Guide, 1st edn. O'Reilly Media Inc., Sebastopol (2015)
5. Gurrin, C., et al.: Introduction to the third annual lifelog search challenge (LSC 2020). In: Gurrin, C., Jónsson, B.Þ., Kando, N., Schöffmann, K., Chen, Y.P., O'Connor, N.E. (eds.) Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, 8–11 June 2020, pp. 584–585. ACM (2020). https://doi.org/10.1145/3372278.3388043
6. Heller, S., et al.: Towards explainable interactive multi-modal video retrieval with Vitrivr. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 435–440. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_41
7. Kratochvíl, M., Veselý, P., Mejzlík, F., Lokoč, J.: SOM-hunter: video browsing with relevance-to-SOM feedback loop. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 790–795. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_71
8. Le, T.K., et al.: Lifeseeker: interactive lifelog search engine at LSC 2019. In: Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC 2019, pp. 37–40. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3326460.3329162
9. Le, T.K., et al.: Lifeseeker 2.0: interactive lifelog search engine at LSC 2020. In: Proceedings of the Third Annual Workshop on Lifelog Search Challenge (2020)
10. LoweDavid, G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)

11. Nguyen, T., Tran-Le, B., Thai, X., Nguyen, T.V., Do, M.N., Tran, M.: Traffic video event retrieval via text query using vehicle appearance and motion attributes. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, 19–25 June 2021, pp. 4165–4172. Computer Vision Foundation/IEEE (2021). https://doi.org/10.1109/CVPRW53098.2021.00470
12. 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., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 760–765. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_66
13. Schoeffmann, K., Lokoc, J., Bailer, W.: 10 years of video browser showdown. In: Chua, T., et al. (eds.) MMAsia 2020: ACM Multimedia Asia, Virtual Event/Singapore, 7–9 March 2021, pp. 73:1–73:3. ACM (2020). https://doi.org/10.1145/3444685.3450215
14. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: an open multilingual graph of general knowledge. In: AAAI (2017)
15. Vo, K., Yamazaki, K., Truong, S., Tran, M., Sugimoto, A., Le, N.: ABN: agent-aware boundary networks for temporal action proposal generation. IEEE Access **9**, 126431–126445 (2021). https://doi.org/10.1109/ACCESS.2021.3110973
16. Vo-Ho, V., Le, N., Yamazaki, K., Sugimoto, A., Tran, M.: Agent-environment network for temporal action proposal generation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, 6–11 June 2021, pp. 2160–2164. IEEE (2021). https://doi.org/10.1109/ICASSP39728.2021.9415101
17. Yuan, J., et al.: Shrec 2018 track: 2D scene sketch-based 3D scene retrieval. In: 3DOR@Eurographics (2018)
18. 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)