Björn Þór Jónsson · Cathal Gurrin · Minh-Triet Tran · Duc-Tien Dang-Nguyen · Anita Min-Chun Hu · Binh Huynh Thi Thanh · Benoit Huet (Eds.)

LNCS 13142

MultiMedia Modeling

28th International Conference, MMM 2022
Phu Quoc, Vietnam, June 6–10, 2022
Proceedings, Part II

2 Part II

28th 2022

Springer

V-FIRST: A Flexible Interactive Retrieval System for Video at VBS 2022

Minh-Triet Tran[1,2,3](✉), Nhat Hoang-Xuan[1,3](✉),
Hoang-Phuc Trang-Trung[1,2,3], Thanh-Cong Le[1,2,3], Mai-Khiem Tran[1,2,3],
Minh-Quan Le[1,3], Tu-Khiem Le[4], Van-Tu Ninh[4], and Cathal Gurrin[4]

[1] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
tmtriet@fit.hcmus.edu.vn, hxnhat18@apcs.fitus.edu.vn,
{tthphuc,ltcong,tmkhiem,lmquan}@selab.hcmus.edu.vn,
[2] John von Neumann Institute, VNU-HCM, Ho Chi Minh City, Vietnam
[3] Vietnam National University, Ho Chi Minh City, Vietnam
[4] Dublin City University, Dublin, Ireland
tukhiem.le4@mail.dcu.ie, tu.ninhvan@adaptcentre.ie, cathal.gurrin@dcu.ie

Abstract. Video retrieval systems have a wide range of applications across multiple domains, therefore the development of user-friendly and efficient systems is necessary. For VBS 2022, we develop a flexible interactive system for video retrieval, namely V-FIRST, that supports two scenarios of usage: query with **text descriptions** and query with **visual examples**. We take advantage of both visual and temporal information from videos to extract concepts related to entities, events, scenes, activities, and motion trajectories for video indexing. Our system supports queries with keywords and sentence descriptions as V-FIRST can evaluate the semantic similarities between visual and textual embedding vectors. V-FIRST also allows users to express queries with visual impressions, such as sketches and 2D spatial maps of dominant colors. We use query expansion, elastic temporal video navigation, and intellisense for hints to further boost the performance of our system.

Keywords: Video retrieval · Interactive system · Sketch retrieval · Color histogram matching · Human-object interaction · Image and video captioning · Moving entity trajectory

1 Introduction

With the rising prevalence of mobile and wearable devices, the amount of image and video data generated has been growing extensively in size. When a user wants to search for some specific items, the idea of browsing through the entire collection quickly becomes infeasible when there are weeks or months worth of data. We note that users often have in mind a picture of what they are searching for. This rough image can be easy to recall, but hard to describe in terms of words, for example, "the picture is mostly blue and green". If they are fortunate

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13142, pp. 562–568, 2022.
https://doi.org/10.1007/978-3-030-98355-0_55

enough, e.g., with the presence of some prominent objects, they can successfully describe the scene with a sentence and use it as a query. However, many times visual information can be difficult to be expressed effectively with words. In those cases, it is helpful to be able to utilize different forms of expressions than textual queries alone.

It would be difficult for users to search for a certain event in a huge amount of video clips. This is a challenging task to create an efficient video retrieval system that can support both accuracy and ease of use. Therefore, the Video Browser (VBS [9]) challenge has been organized for many years to encourage researchers worldwide to develop and enhance retrieval systems for video clips.

For VBS 2022, we propose and develop V-FIRST, a **F**lexible **I**nteractive **R**etrieval **S**ys**T**em for **V**ideos. Our system allows the user to specify various visual cues such as the global color histogram of the image, local prominent color patches, and a sketch of the sought scene. As visual information does not have a simple expression such as a sentence or clause, our system is made up of many retrieval modalities that utilize different visual aspects of the videos, and the user may choose to specify some, or all of them to assist with retrieval. With sufficient usage and feedback, over time, we can evaluate the compatibility of each modality with each specific use case and give recommendations to users based on that.

Our system is based on FIRST [13,15], which was a system used for retrieving images from lifelog, however, we have carefully extended it to accommodate for video retrieval. The first notable addition is temporal information, which was not present in images. Our system supports activity recognition and action proposal [16,17]. Second, our system enhances the captioning ability of images by utilizing optical character recognition (OCR) and extending the set of defined concepts. This enables better semantic similarity between the embeddings of the (textual) query, the generated caption, and the image itself. Third, apart from descriptive language, we support the inclusion of visual impressions, such as global color histogram, prominent color patches, and sketches of the video. Finally, we aim to integrate intellisense, a simple context-based suggestion system that asks the user some questions that would quickly reduce the set of possible answers to the query. This allows our system to actively assist the user in the retrieval process, instead of being passive and completely relying on the user's ability to generate effective queries.

We also implement query expansion in our system based on visual similarity/dissimilarity by allowing the user to specify scenes that are close to the sought scene or are irrelevant. Those samples then act as positive/negative samples and can be used as an input for the next query. Furthermore, our system allows temporal exploration of the retrieved results in both directions, as a means for the user to continue searching in the time dimension or to simply verify the correctness of the retrieved videos.

2 Related Work

The Video Browser Showdown (VBS [9]) competition serves as a benchmark of interactive video retrieval systems that are held annually. In the 2021 edition, vitrivr [3] managed to achieve the best score. This system is an example of a multimodal system, in which they allow users to search using pre-defined concepts and a rough color sketch of the scene. Other advanced features include temporal search, which is searching for concepts appearing in a specific order, and semantic sketching, where users essentially input a sketch of the scene's semantic segmentation result. Notably, all of these input methods are based on the same list of pre-defined concepts.

Traditional methods implement some subset of the features mentioned and attempt to enhance them in different ways. IVOS [7] and VISIONE [1] both enhance their retrieval performance by allowing users to specify the location of objects, in addition to normal queries. SOMHunter [4] upgrades object localization to concept localization by allowing clauses (e.g., woman eating) to be localized. Also, both SOMHunter [4] and VISIONE [1] supports temporal query by allowing users to input two queries instead of one.

VideoGraph [8] deviates by using Wikipedia to generate a knowledge graph that allows the user to search using semantically related concepts, e.g., they can search for "engine" and find scenes with cars and motorbikes. For input methods, aside from typical keyboard and mouse, vitrivr-VR [11] and EOLAS [12] utilize Virtual Reality to allow users to have an immersive view of the results. These systems use interaction within the virtual environment as a replacement for complex text queries.

3 V-FIRST- A Flexible Interactive Retrieval System for Video

3.1 System Overview

Figure 1 shows the overview structure of our proposed system, V-FIRST, which is developed from our current lifelog retrieval systems FIRST [13,15]. Thanks to the flexible architecture of FIRST, we can easily enhance and integrate new features for video clip retrieval into our system.

We develop two sets of query processing modules to support two types of query scenarios with text descriptions and visual examples. For queries with text descriptions, our system allows users to search with visual concepts, activities, motion of entities, and with free-text descriptions (see Sect. 3.2). For queries with visual examples, we support searching with sketches and spatial maps of colors (see Sect. 3.3).

We also develop additional functions to assist users. With Query Expansion, the system can use a result candidate as a positive or negative example for result refinement. We also assist users in quickly exploring a video clip with a flexible timeline to verify past or future events from an initial moment. We also develop a preliminary intellisense function to provide context-based hints to users to quickly refine the result. Finally, our system also allows users to combine different functions to perform a complex search or refine their search results.

3.2 Query Processing with Text Descriptions

For each video clip, we extract visual concepts [2,10] related entities appearing in each video shot, an equal-sized sequence of frames. We also capture the scene attributes and categories [18], and scene texts [6] from a video shot for indexing.

Besides useful information from each still image, we also take advantage of temporal information from video clips. We use action proposal detection [16,17] to find potential meaningful activities in a video clip, then classify these proposals for known activity categories.

Object movement can be useful information for retrieval. We develop the idea by Nguyen et al. [5,6] for traffic video event retrieval from text description into our system for general video cases. We evaluate the similarity between the trajectories of main entities in a video clip with the text description to find appropriate video clips having a certain event about object movement.

To further bridge the gap between text and visual information, we generate captions for main shots in a video clip so that we can evaluate the similarity between an input text query phrase with generated captions. Based on the simple idea of utilizing multiple input branches, including OCR features and object features, into transformer architecture and using a copy mechanism to generate captions, we implement a graph neural network to learn the relationship between OCR features and object features as another input branch. In this way, we successfully integrate a useful captioning module for our solution that can exploit spatial relationships between objects and scene texts.

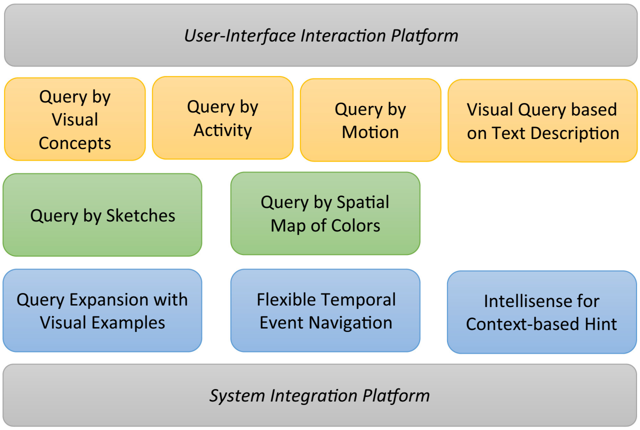
User-Interface Interaction Platform
Query by Visual Concepts
Query by Activity
Query by Motion
Visual Query based on Text Description
Query by Sketches
Query by Spatial Map of Colors
Query Expansion with Visual Examples
Flexible Temporal Event Navigation
Intellisense for Context-based Hint
System Integration Platform

Fig. 1. System overview of our query system.

To provide users with a natural way to search with free text, we employ our Self-Attention based Joint Embedding Model (SAJEM [14]) in our retrieval system. In this way, our system encodes both a text query and a photo/sequence of photos into an embedding space to measure their similarity.

3.3 Query Processing with Visual Examples

In the Known-Item Search setting of VBS, we are given a dataset of video clips. We are presented with a query video, and our task is to use our system to find that clip. To express which data we want to retrieve, we need a query. Perhaps the most popular form of one is a textual sentence, which is used by Google, the most popular search engine. This form of expression is useful if our information need is simple enough to be parsed into a short sentence. For example, we might describe a picture in terms of the main objects and their interactions. However, what if the picture cannot be well described using words?

We can build a text-based system, and try to describe the query clips one by one. Another approach is searching based on visual cues (e.g., colors). This visual-based approach is content-agnostic and has the potential to outperform text-based approaches in certain circumstances. It is also more complicated to implement, since, unlike language, there are many ways to represent visual cues, and finding an efficient scheme to use and to search is challenging.

With a visual example, instead of describing that example in text and reusing the existing solutions in Sect. 3.2, we develop some preliminary utilities to assist users in describing the visual impressions from the example, such as global color histogram, prominent color patches, and sketches of the video.

We represent the query as a $m \times n$ grid, where each cell contains the main color. When the grid is overlaid on images being searched, the color of each cell represents the (expected) dominant color in that cell. This way, the user can specify the dominant color of each part of the image, instead of the whole image. Even better, they can choose to only identify parts that have a relatively uniform color, the rest can be left as "unspecified".

We also allow users to quickly sketch the main visual features that they remember most from the visual examples. Of course, this way requires the user to have good skills in arts. We also divide an image into a grid of cells, and we estimate the histogram of gradient orientation in each cell from the sketch. We use asymmetric distance to compare the fitness of the input sketch query with frames in the video to find candidates for the query.

4 Conclusion

In this paper, we introduce our solution for retrieving video clips in a large collection of videos. Our goal is to provide an easy-to-use system that can assist users in expressing their query needs in a flexible way. Therefore we provide two sets of utilities to handle queries in text description mode and queries with visual examples. These two scenarios correspond to the Ad-hoc Video Search (AVS) and Known-Item Search (KIS) of the VBS challenge.

We take advantage of the flexible architecture of our retrieval system to integrate different components for query processing. For a query with text description, users can search with visual concepts related to entities, scene attributes and categories, activities, scene-text, and even in free-text format. For a query with visual examples, users can define a draft 2D map of dominant colors or sketch out main visual features to find candidate images from the dataset. Query Expansion and Flexible Video Navigation assist users in further exploring potential results from an initial candidate. We also integrate a preliminary intellisense feature to automatically ask users questions to quickly refine the candidate list.

Acknowledgement. The team would like to thank Vinh-Hung Tran, Trong-Thang Pham for the enhanced captioning module; Trong-Tung Nguyen and Huu-Nghia Nguyen-Ho for the human-object interaction module; Tien-Phat Nguyen and Ba-Thinh Tran-Le for the moving trajectory retrieval method.

Hoang-Phuc Trang-Trung, Thanh-Cong Le, and Mai-Khiem Tran were funded by Vingroup Joint Stock Company and supported by the Domestic Master/ PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA), code VINIF.2020.ThS.JVN.03, VINIF.2020. ThS.JVN.05, and VINIF.2020.ThS.JVN.06, respectively.

The work was funded by Gia Lam Urban Development and Investment Company Limited, Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

References

1. Amato, G., et al.: VISIONE at video browser showdown 2021. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 473–478. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_47
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
3. 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
4. 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
5. Liu, Y., Chen, H., Shen, C., He, T., Jin, L., Wang, L.: AbcNet: real-time scene text spotting with adaptive Bezier-curve network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020
6. Nguyen, N., et al.: Dictionary-guided scene text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7383–7392, June 2021
7. Ressmann, A., Schoeffmann, K.: IVOS - the ITEC interactive video object search system at VBS2021. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 479–483. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_48

8. Rossetto, L., et al.: VideoGraph – towards using knowledge graphs for interactive video retrieval. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 417–422. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_38
9. 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)
10. Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10778–10787 (2020)
11. Tran, D., et al.: A VR interface for browsing visual spaces at VBS2021, pp. 490–495 (2021)
12. Tran, L.-D., et al.: A VR interface for browsing visual spaces at VBS2021. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 490–495. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_50
13. Tran, M., et al.: FIRST - flexible interactive retrieval system for visual lifelog exploration at LSC 2020. In: Gurrin, C., et al. (eds.) Proceedings of the Third ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2020, Dublin, Ireland, 8–11 June 2020, pp. 67–72. ACM (2020)
14. Trang-Trung, H., Le, H., Tran, M.: Lifelog moment retrieval with self-attention based joint embedding model. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 22–25 September 2020. CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020). http://ceur-ws.org/Vol-2696/paper_60.pdf
15. Trang-Trung, H., et al.: Flexible interactive retrieval system 2.0 for visual lifelog exploration at LSC 2021. In: Gurrin, C., et al. (eds.) Proceedings of the 4th Annual on Lifelog Search Challenge, LSC@ICMR 2021, Taipei, Taiwan, 21 August 2021, pp. 81–87. ACM (2021)
16. 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)
17. 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)
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**, 1452–1464 (2017)