Will VISIONE Remain Competitive in Lifelog Image Search?

Giuseppe Amato
giuseppe.amato@isti.cnr.it
CNR-ISTI
Pisa, Italy

Paolo Bolettieri
paolo.bolettieri@isti.cnr.it
CNR-ISTI
Pisa, Italy

Fabio Carrara
fabio.carrara@isti.cnr.it
CNR-ISTI
Pisa, Italy

Fabrizio Falchi
fabbrizio.falchi@isti.cnr.it
CNR-ISTI
Pisa, Italy

Claudio Gennaro
claudio.gennaro@isti.cnr.it
CNR-ISTI
Pisa, Italy

Nicola Messina
nicola.messina@isti.cnr.it
CNR-ISTI
Pisa, Italy

Lucia Vadicamo*
lucia.vadicamo@isti.cnr.it
CNR-ISTI
Pisa, Italy

Claudio Vairo
claudio.vairo@isti.cnr.it
CNR-ISTI
Pisa, Italy

ABSTRACT

VISIONE is a versatile video retrieval system supporting diverse search functionalities, including free-text, similarity, and temporal searches. Its recent success in securing first place in the 2024 Video Browser Showdown (VBS) highlights its effectiveness. Originally designed for analyzing, indexing, and searching diverse video content, VISIONE can also be adapted to images from lifelog cameras thanks to its reliance on frame-based representations and retrieval mechanisms.

In this paper, we present an overview of VISIONE's core characteristics and the adjustments made to accommodate lifelog images. These adjustments primarily focus on enhancing result visualization within the GUI, such as grouping images by date or hour to align with lifelog dataset imagery. It's important to note that while the GUI has been updated, the core search engine and visual content analysis components remain unchanged from the version presented at VBS 2024. Specifically, metadata such as local time, GPS coordinates, and concepts associated with images are not indexed or utilized in the system. Instead, the system relies solely on the visual content of the images, with date and time information extracted from their filenames, which are utilized exclusively within the GUI for visualization purposes.

Our objective is to evaluate the system's performance within the Lifelog Search Challenge, emphasizing reliance on visual content analysis without additional metadata.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Users and interactive retrieval**; **Retrieval models and ranking**; *Search engine architectures and scalability*; **Multimedia and multimodal retrieval**; **Video search**.

LSC '24, June 10, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0550-2/24/06.
https://doi.org/10.1145/3643489.3661122

KEYWORDS

multimedia retrieval, image search, cross-modal search, interactive system, lifelong image, egocentric images

ACM Reference Format:
Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2024. Will VISIONE Remain Competitive in Lifelog Image Search?. In *The 7th Annual ACM Lifelog Search Challenge (LSC '24), June 10, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3643489.3661122

1 INTRODUCTION

In recent years, the proliferation of wearable sensors has paved the way for lifelogging to emerge as a potentially widespread activity. Lifeloggers, equipped with wearable devices, can capture rich streams of audio-visual records, geographic locations, and biometric data. This continuous and comprehensive recording of an individual's daily experiences yields a wealth of multimodal lifelog data, which holds significant potential for various applications, ranging from personal memory augmentation to healthcare monitoring and beyond.

The abundance and diversity of lifelog data offer researchers and practitioners considerable opportunities for analysis and exploration. However, efficient retrieval and analysis of multimodal lifelog data pose significant challenges. In this context, the Lifelog Search Challenge (LCS) [11, 17] has emerged as a notable initiative for advancing research and innovation in lifelog data retrieval. Since its first edition in 2018, LSC has served as an interactive competition in which teams compete with each other to develop the leading PoV Egocentric Lifelog retrieval tool. LSC'24 [10] builds upon the dataset utilized in LSC'23/22, which is an extensive 18-month multimodal lifelog collection gathered by a single lifelogger. This dataset comprises over 725,000 fully redacted and anonymized wearable camera images captured with a Narrative Clip device at a resolution of 1024 x 768 between 2019 and 2020. The images are organized sequentially and aligned with UTC time, supplemented by minute-by-minute textual metadata containing time and location details, music listening history, biometrics, semantic names for locations, flight details, and other information pre-extracted from the data.

This paper presents the VISIONE video retrieval system, demonstrating its applicability for lifelog image search. VISIONE [1–3] is an interactive video search tool that uses state-of-the-art Artificial Intelligence techniques for visual content analysis and advanced indexing techniques for scalability. While VISIONE has not previously participated in the LSC or engaged with egocentric image data, it has achieved significant success in the Video Browser Showdown (VBS) [13], an international competition for interactive video search. In the last editions of VBS, held on January 29, 2024, VISIONE emerged as the top interactive video search system, securing first place in four out of seven tasks evaluated by both expert and novice users[1]. Given its good performance recently exhibited in retrieving diverse video content, we are interested in assessing its performance in lifelong image retrieval without substantial modifications to its core search system. For this purpose, VISIONE aims to participate in LCS'24 with its existing implementation. It's worth noting that while LSC provides additional metadata, such as local time and location information, we did not incorporate this data into our index. As a result, our system does not currently support specifying time or location constraints during search queries. However, we have processed the LSC dataset to align with VISIONE requirements and have updated certain aspects of the user interface to facilitate searches on the LSC data, as described in the following sections.

The remainder of this paper is structured as follows: Section 2 provides an overview of the VISIONE system; Section 3 delineates the adaptations implemented in the system to facilitate LSC data search, presents qualitative results on some LSC'23 tasks, and discusses system limitations; Section 4 provides concluding remarks.

2 VISIONE SYSTEM OVERVIEW

The VISIONE retrieval system supports four main search functionalities:

- *Text-to-Image search*, where users can use textual descriptions in natural language to search for a target image.
- *Similarity search* where users can use an image displayed in the browsing interface, or uploaded from a file/URL, to search for similar content.
- *Object and Color based search*, where users can specify the positions of objects and colors appearing in a scene of interest on a canvas.
- *Temporal search*, where users can specify two distinct queries (textual or object/color-based) to search for two temporally close images in a video.

To support effective free-text searches, we employ three multimodal feature extractors, each based on a pre-trained model. Specifically, we utilize the OpenCLIP ViT-L/14 model [2] [16], named ClipLAION in the GUI, which is pre-trained on the LAION-2B dataset. Additionally, we incorporate the CLIP2Video [3] [8] and ALADIN[4] [14] models to further augment our search capabilities. By default, our system combines all three models for free-text search, leveraging a late fusion approach to merge the results. However, users can select a specific model via the dedicated radio button provided in the user interface (see Figure 1a).

The ALADIN and CLIP2Video features are also used to support semantic similarity search when an image example is provided as a query. Additionally, the visual similarity search utilizes features extracted using DINOv2 [15], which have demonstrated effectiveness across various image-level visual tasks, including image classification, instance retrieval, and video understanding. By default, our system employs a late fusion of the similarity search results obtained using DINOv2, ALADIN, and CLIP2Video. However, our GUI offers an "Advanced mode" accessible through a button located in the top-left corner of the interface (Figure 1a). This mode gives users finer control over the model selection for similarity search (Figure 1f).

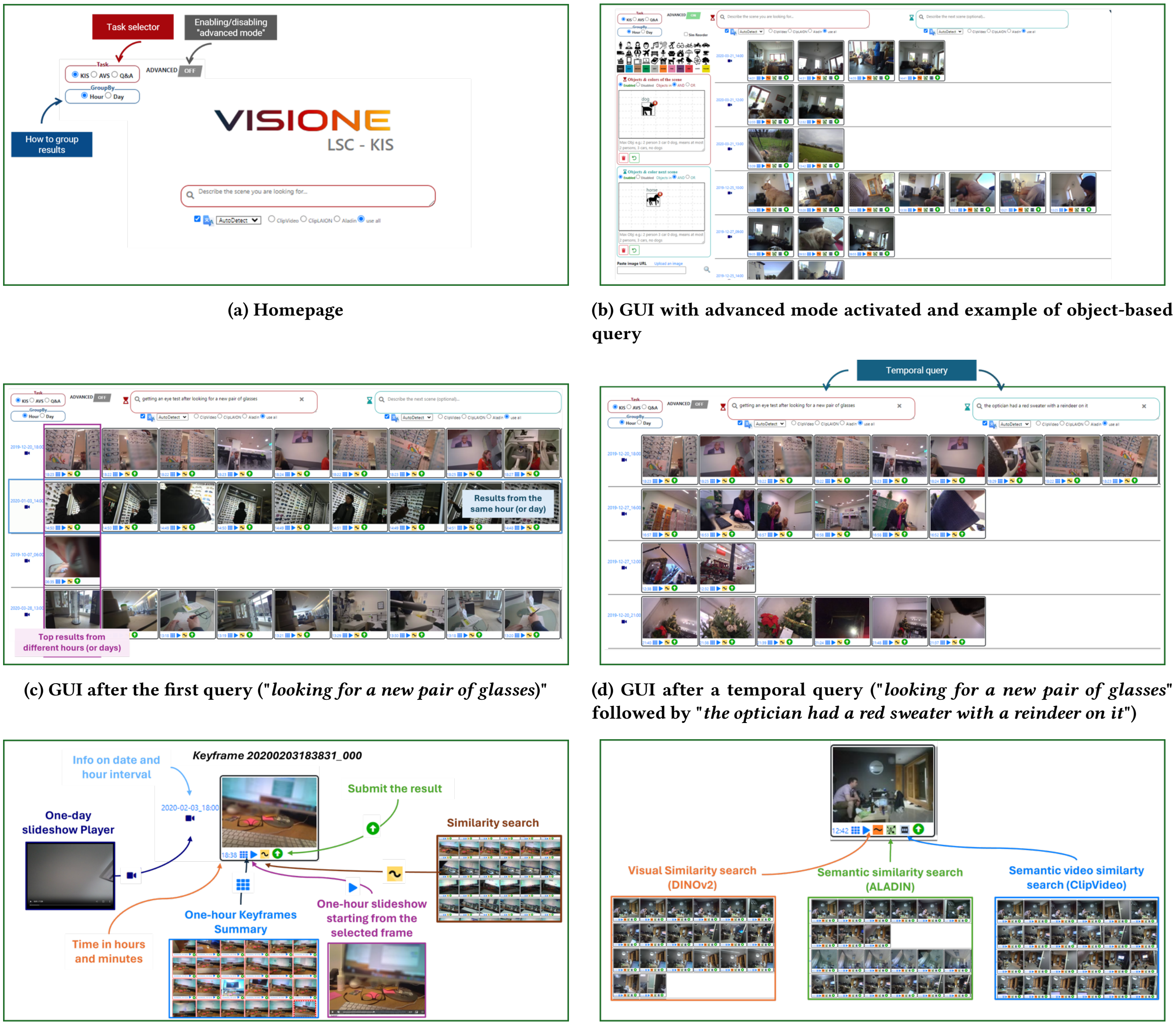
In the advanced mode, users can also perform queries based on objects and colors by dragging or inserting such objects/colors into dedicated GUI canvases (Figure 1b). Object-based queries rely on annotations generated using the following models: VfNet [19], Mask R-CNN [12], and Faster R-CNN [9] trained on COCO, LVIS, and Open Images V4 datasets, respectively. Additionally, color annotation is conducted using two chip-based color naming techniques [5, 18].

The VISIONE system also supports temporal queries where the users can specify two separate queries that describe scenes occurring close in time within a specific video segment. Employing a temporal quantization and matching technique, the system searches for videos containing images that fulfill the description outlined in both the initial and subsequent queries. The second text box in the right part of the GUI (Figure 1d) can be used to issue the second query.

For our indexing and searching strategy, we utilize two different access methods. Firstly, the Facebook FAISS library[5] is employed to store and access both CLIP2Video and ClipLAION features. Secondly, a separate index built using Apache Lucene [6] is specifically tailored to store all other descriptors. Notably, when indexing all the extracted descriptors with Lucene, we have developed specialized text encodings based on the Surrogate Text Representations (STRs) approach, as described in previous works [1, 4, 6, 7].

The results are presented to the user grouped by rows, with the most relevant search results presented in the first column (Figure 1c). While the standard grouping in VISIONE is by video, support is provided for grouping by hour and day for LSC data, as described in the following section. By right-clicking on an image, users can preview images captured just before or after the selected image. Left-clicking opens a larger view of the image at the top of the browsing interface, allowing users to navigate to previous or next videos using the keyboard arrows. Each result entry in the GUI includes several interactive elements (see Figure 1e): keyframe ID (displayed as time for LSC data), which links to the full-size keyframe and provides a summary of objects extracted from it; a grid button, which opens a window displaying all indexed images of a video (images acquired at the same hour of the selected image for LSC

[1] https://videobrowsershowdown.org/hall-of-fame/
[2] https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K
[3] https://github.com/CryhanFang/CLIP2Video
[4] https://github.com/mesnico/ALADIN
[5] https://github.com/facebookresearch/faiss
[6] https://lucene.apache.org/


Task selector
Enabling/disabling "advanced mode"
How to group results
Task
KIS AVS Q&A
GroupBy
Hour Day
ADVANCED OFF
VISIONE
LSC - KIS
Describe the scene you are looking for...
AutoDetect
ClipVideo ClipLAION Aladin use all
Temporal query
getting an eye test after looking for a new pair of glasses
the optician had a red sweater with a reindeer on it
Results from the same hour (or day)
Top results from different hours (or days)
Info on date and hour interval
Keyframe 20200203183831_000
2020-02-03_18:00
Submit the result
One-day slideshow Player
Similarity search
One-hour Keyframes Summary
One-hour slideshow starting from the selected frame
Time in hours and minutes
Visual Similarity search (DINOv2)
Semantic similarity search (ALADIN)
Semantic video similarity search (ClipVideo)

(a) Homepage

(b) GUI with advanced mode activated and example of object-based query

(c) GUI after the first query ("*looking for a new pair of glasses*")

(d) GUI after a temporal query ("*looking for a new pair of glasses*" followed by "*the optician had a red sweater with a reindeer on it*")

(e) Visualizations and Browsing options (advanced mode OFF)

(f) Similarity search options (advanced mode ON)

Figure 1: VISIONE's GUI

data); the play button, initiating video playback from the selected frame (a slideshow of images captured within the same hour for LSC); and the green button used for submitting results during the competition.

For further details on the system's architecture and its user interface, please refer to our previous publications, which describe these details [1–3].

3 VISIONE AT LIFELOG SEARCH CHALLENGE

How will VISIONE perform at LSC'24 without changing its core functionalities? We have neither integrated new models for feature extraction from images nor added LSC metadata (e.g., information about time or space) to our index. Instead, our focus has been solely on arranging the LSC dataset according to VISIONE requirements and implementing some specific changes to the GUI to enhance search functionality within this lifelog imagery. We describe all the changes made to VISIONE in the following.

Data preprocessing and analysis. VISIONE is a system originally designed for video search rather than image search. Although the system typically expects a set of videos as input, it is important to note that all content analyses, except for the Clip2Video model, are performed at the keyframe level. Even the search modules within the system are frame-based. Therefore, the core functionality of VISIONE remains applicable to image datasets without requiring modification. However, there are a few functionalities, including extraction of Clip2Video features and some visualization functionalities in the GUI, which expect video input. The LCS dataset consists of images rather than videos; however, it can be conceptualized as a series of videos spanning entire hours (or other time units). To accommodate this dataset, we generated videos by compiling the images corresponding to each hour into slideshows, effectively encapsulating one hour of activity in each video. These hour-based videos were utilized as input for content analysis and indexing within the VISIONE system. We also provided the pre-extracted keyframes, corresponding to the original LSC images, as input to the system, preventing using a scene detector. This approach allowed us to leverage the VISIONE system to process the LCS data and conduct all analyses, including those involving the Clip2Video model. Therefore, search results in VISIONE return the original images from the LCS dataset, with associated videos comprising one-hour slideshows.

Visualization of time information. Although we have not indexed the metadata associated with the LCS images, it's worth noting that the file names of the images contain information about the date, hour, minutes, and seconds when they were captured. For example, an image named "20200203183831_000.png" was acquired on February 3, 2020, at 18:38 and 31 seconds, according to the UCT time zone set in the acquisition device. To enhance the user experience while browsing the results, we have made modifications to our GUI. Instead of displaying the original identifiers of each image and video, we present the time information in a more readable format. Specifically, only the time is displayed in the HH:MM format for each image. For video identifiers, we present the date and hour in the format YYYY-MM-DD_HH:00. For instance, "2020-02-03_18:00" indicates the video spanning from 6:00 PM to 6:59 PM on February 3, 2020. For an illustration, refer to Figure 1e.

Full-day video. Since the videos associated with each image correspond to a single hour of the day, we have introduced the option to view a video of the entire day within the interface. To achieve this, we have (1) generated full-day videos as a simple slideshow of the images captured throughout the day, (2) added a dedicated button placed below the video identifier in the left part of the browsing interfaces (see Figure 1c) that initiates the playback of the full-day video. This feature allows the user to quickly review the events of a whole day, even if the search results only display an hour of video (due to the specific query referring to that moment). This is very useful in certain tasks where the user needs to have a look at an entire day rather than at a single moment of the day.

Temporal query. In LCS, tasks may involve situations of different parts of the day. For example, one hint might refer to the morning, while another to the evening. Consequently, we have adjusted the behavior of the temporal query to better accommodate these scenarios. By default, in VISIONE, the temporal search conducts two independent searches and matches the results within a time interval of 21 seconds (video time). However, in this case, we have updated the temporal query to utilize the date information contained in the image identifiers rather than the time at which they appear in the video. This adjustment enables the matching of results captured on the same day. This change is the only modification made to the core part of VISIONE, and it is relatively simple and lightweight. It only involves adjusting a criterion used to compare and aggregate the results of two independent queries.

Group by day. To enhance result browsing, we have introduced a radio button, placed just below the task-switching button, which allows users to choose how the results are displayed (see Figure 1a). By default, results are grouped by video, with frames from a single one-hour video presented in each row. Alternatively, the "group-by-day" option organizes results by day, displaying keyframes from the most relevant one-hour videos spanning the entire day. This feature aids in tasks involving hints from different moments of the day.

3.1 Qualitative results on LSC'23 tasks

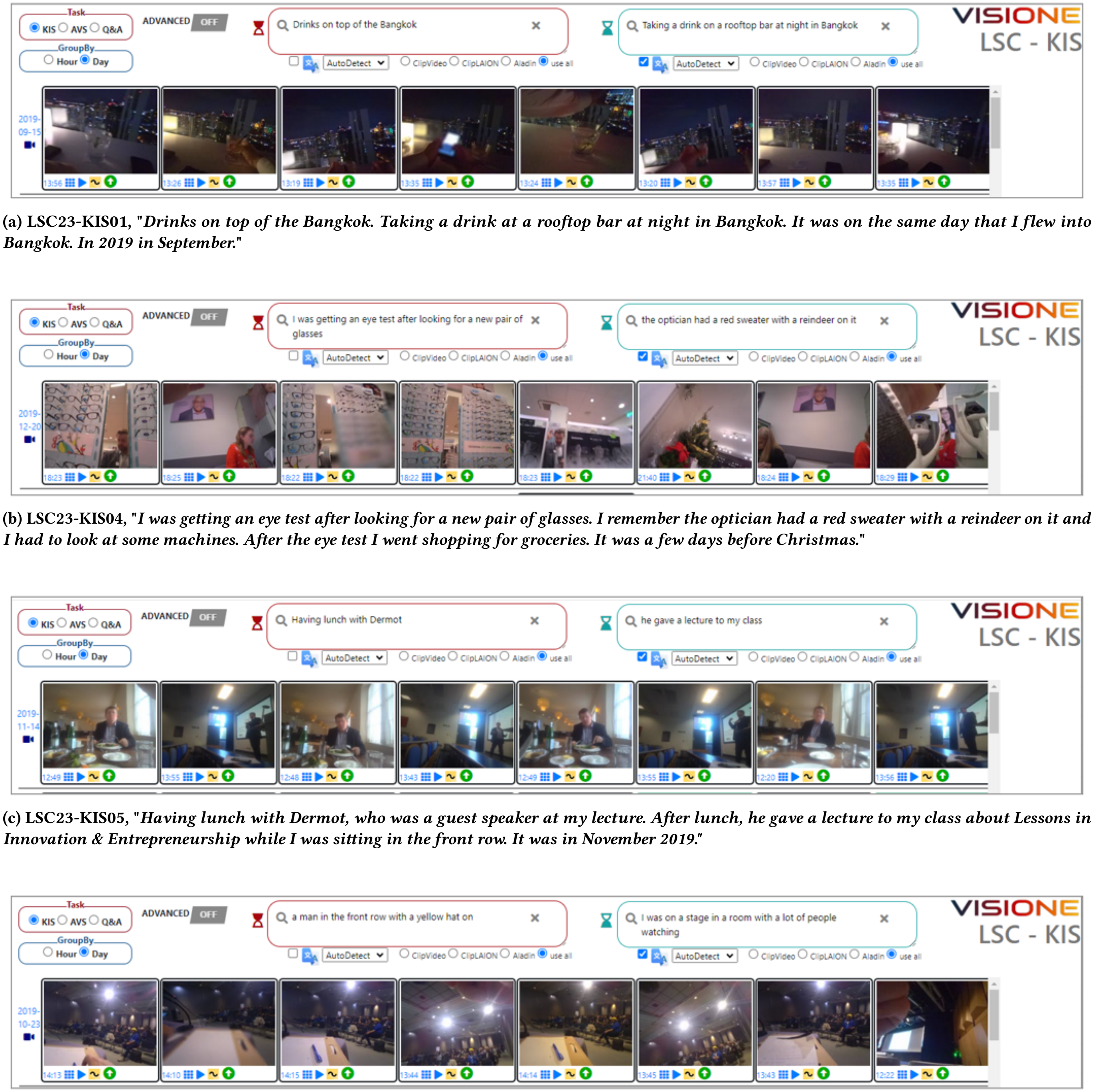
The tasks proposed in the 2023 edition of the LSC competition are available on the LSC webpage[7] and can be leveraged in the development and testing of systems intending to participate in LSC'24. In particular, the competition comprised 10 question/answering (QA) tasks, 10 ad-hoc search (AD) tasks, and 10 textual known-item search (KIS) tasks.

We utilized these tasks to get initial qualitative feedback on the performance of VISIONE for lifelog image retrieval. The results obtained were encouraging, particularly for KIS tasks. Specifically, with a single query or with two queries combined with temporal search, we were able to identify the correct result in the first position in 8 out of 10 KIS tasks. Examples are illustrated in Figure 2. However, in the remaining two tasks, we could not find the correct solution.

In our tests, we attempted to retain as many original task sentences as possible. Surprisingly, reworking the text to what we believed would be better suited to our system often resulted in worse outcomes. For instance, in the task LSC23-KIS05 (see Figure 2c), employing temporal search between the original task sentences *"Having lunch with Dermot"* and "*he gave a lecture to my class"* returned the correct result in the first position. Conversely, using a more generic phrasing for temporal searches, such as *"Having lunch with a friend"* and *"a person giving a lecture to a class"*, yielded inferior results, with the correct moment appearing farther down the results list. The employed models lack knowledge of specific identities. However, during training, they may have encountered images with named captions (featuring, for example, famous people). Thus, specifying names might enhance the model's ability to discern situations. For instance, the network has biases and may have learned that certain names are associated with particular genders and ethnicities. Indeed, we have observed significant variations in the results by querying with "Having lunch with [NAME]," varying the name among some typical names from different regions worldwide.

Another observation we made is that QA-type tasks pose greater challenges for our system, particularly when they heavily rely on temporal and georeferenced information that is not included in our index. For instance, tasks such as LSC23-QA06 *"On what date*

[7] http://lifelogsearch.org/lsc/resources/LSC23-Topics.txt



Task: ● KIS ○ AVS ○ Q&A — ADVANCED OFF — GroupBy: ○ Hour ● Day

Drinks on top of the Bangkok — □ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

Taking a drink on a rooftop bar at night in Bangkok — ☑ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

(a) LSC23-KIS01, "*Drinks on top of the Bangkok. Taking a drink at a rooftop bar at night in Bangkok. It was on the same day that I flew into Bangkok. In 2019 in September.*"

Task: ● KIS ○ AVS ○ Q&A — ADVANCED OFF — GroupBy: ○ Hour ● Day

I was getting an eye test after looking for a new pair of glasses — □ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

the optician had a red sweater with a reindeer on it — ☑ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

(b) LSC23-KIS04, "*I was getting an eye test after looking for a new pair of glasses. I remember the optician had a red sweater with a reindeer on it and I had to look at some machines. After the eye test I went shopping for groceries. It was a few days before Christmas.*"

Task: ● KIS ○ AVS ○ Q&A — ADVANCED OFF — GroupBy: ○ Hour ● Day

Having lunch with Dermot — □ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

he gave a lecture to my class — ☑ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

(c) LSC23-KIS05, "*Having lunch with Dermot, who was a guest speaker at my lecture. After lunch, he gave a lecture to my class about Lessons in Innovation & Entrepreneurship while I was sitting in the front row. It was in November 2019.*"

Task: ● KIS ○ AVS ○ Q&A — ADVANCED OFF — GroupBy: ○ Hour ● Day

a man in the front row with a yellow hat on — □ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

I was on a stage in a room with a lot of people watching — ☑ AutoDetect ○ ClipVideo ○ ClipLAION ○ Aladin ● use all

(d) LSC23-KIS08, "*There was a man in the front row with a yellow hat on. I was on a stage in a room with a lot of people watching. I was on some sort of panel and writing notes on paper. I remember the man had a blue sweater/top on also. It was in France at ACM MM2019.*"

Figure 2: Examples of KIS tasks where the correct results were found in the first position with a single temporal query.

in 2019 did I go homewares shopping around midnight in Ireland?" are difficult to solve and require considerable user attention while browsing the results. Despite user intervention, we were not always able to find the correct answer for the task.

3.2 Limitations

The fact that we have not processed the metadata associated with the dataset poses significant limitations in searching, particularly concerning specific dates and locations. However, we mitigate this issue to some extent by using effective large multimodal models. For instance, one of the models employed, the CLIP ViT L/14 [8] [16], was trained on the large-scale LAION-2B dataset. Consequently, it is capable, in some instances, of discerning location-related information. For example, as demonstrated in Figure 2a, it retrieved Bangkok-related results without any location-associated information in the index.

Additionally, the multimodal models used in VISIONE may effectively distinguish between day and night in outdoor images. However, this capability is still limited, as the system does not support specific filters on time, date, day of the week, etc. One mitigation to this limitation is that our GUI is a web-based interface, allowing users to utilize the browser's find functionality to highlight results from a specific date (for example, all videos from July 2019). Although this does not filter out results, it can aid users in focusing only on interesting results while browsing.

Some of the images in the LSC dataset have been captured from various locations worldwide, and the local time information about when the shots were taken is stored in the metadata associated with the dataset. However, the filename, which contains the date and time of the image, is still set to UCT time. Since we did not process any metadata, we cannot determine the local time of the images if they are from countries outside the UCT time zone. Therefore, if a task refers to moments of the day based on shots taken with a different timestamp, the user cannot rely on the temporal information in the filename, as it does not match the actual time of that image.

4 CONCLUSIONS

In this paper, we provide an overview of the VISIONE video retrieval system and we describe the adjustments made to accommodate lifelog images. Despite relying solely on visual content analysis without additional metadata (e.g., local time, GPS coordinates, semantic locations, etc), preliminary results suggest that the system remains competitive for retrieving lifelong images, particularly for KIS tasks. However, it may exhibit limited performance in QA, or in tasks that heavily rely on temporal and georeferenced information. As a next step, we aim to evaluate its performance in the upcoming Lifelog Search Challenge.

ACKNOWLEDGMENTS

This work was partially funded by AI4Media - A European Excellence Centre for Media, Society, and Democracy (EC, H2020 n. 951911), MUCES - a MUltimedia platform for Content Enrichment and Search in audiovisual archive project (EU NextGenerationEU PNRR, CUP B53D23026090001), National Centre for HPC, Big Data and Quantum Computing – HPC (EU NextGenerationEU PNRR, CUP B93C22000620006), and FAIR – Future Artificial Intelligence Research - Spoke 1 (EU NextGenerationEU PNRR M4C2 PE00000013). This manuscript reflects only the authors' views and opinions; neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. 2021. The VISIONE video search system: exploiting off-the-shelf text search engines for large-scale video retrieval. *Journal of Imaging* 7, 5 (2021), 76.

[2] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2023. VISIONE at Video Browser Showdown 2023. In *MultiMedia Modeling*. Springer, 615–621.

[3] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2024. VISIONE 5.0: Enhanced User Interface and AI Models for VBS2024. In *International Conference on Multimedia Modeling*. Springer, 332–339.

[4] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. 2020. Large-scale instance-level image retrieval. *Information Processing & Management* 57, 6 (2020), 102100.

[5] Robert Benavente, Maria Vanrell, and Ramon Baldrich. 2008. Parametric fuzzy sets for automatic color naming. *JOSA A* 25, 10 (2008), 2582–2593.

[6] Fabio Carrara, Claudio Gennaro, Lucia Vadicamo, and Giuseppe Amato. 2023. Vec2Doc: Transforming Dense Vectors into Sparse Representations for Efficient Information Retrieval. In *Similarity Search and Applications*. Springer, Cham.

[7] Fabio Carrara, Lucia Vadicamo, Claudio Gennaro, and Giuseppe Amato. 2022. Approximate Nearest Neighbor Search on Standard Search Engines. In *Similarity Search and Applications*, Tomáš Skopal, Fabrizio Falchi, Jakub Lokoč, Maria Luisa Sapino, Ilaria Bartolini, and Marco Patella (Eds.). Springer International Publishing, Cham, 214–221.

[8] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097* (2021).

[9] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[10] Cathal Gurrin, Graham Healy, Liting Zhou, Björn Þór Jónsson, Duc Tien Dang Nguyen, Jakub Lokoc, Luca Rossetto, Minh-Triet Tran, Steve Hodges, Werner Bailer, and Klaus Schoeffmann. 2024. Introduction to the Seventh Annual Lifelog Search Challenge, LSC'24. *International Conference on Multimedia Retrieval (ICMR'24)*. https://doi.org/10.1145/3652583.3658891

[11] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *International Conference on Multimedia Retrieval (ICMR'22). ACM.*

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[13] Jakub Lokoč, Stelios Andreadis, Werner Bailer, Aaron Duane, Cathal Gurrin, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peška, Luca Rossetto, Loris Sauter, Konstantin Schall, Klaus Schoeffmann, Omar Shahbaz Khan, Florian Spiess, Lucia Vadicamo, and Stefanos Vrochidis. 2023. Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th VBS. *Multimedia Systems* (24 Aug 2023). https://doi.org/10.1007/s00530-023-01143-5

[14] Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. 2022. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. *arXiv preprint arXiv:2207.14757* (2022).

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[17] Ly-Duyen Tran, Manh-Duy Nguyen, Duc-Tien Dang-Nguyen, Silvan Heller, Florian Spiess, Jakub Lokoč, Ladislav Peška, Thao-Nhu Nguyen, Omar Shahbaz Khan, Aaron Duane, et al. 2023. Comparing Interactive Retrieval Approaches at the Lifelog Search Challenge 2021. *IEEE Access* (2023).

[18] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. 2009. Learning color names for real-world applications. *IEEE Transactions on Image Processing* 18, 7 (2009), 1512–1523.

[19] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. 2021. VarifocalNet: An IoU-aware Dense Object Detector. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[8] https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K