Flexible Interactive Retrieval SysTem 3.0 for Visual Lifelog Exploration at LSC 2022

Nhat Hoang-Xuan
University of Science
Vietnam National University
Ho Chi Minh city, Vietnam

Hoang-Phuc Trang-Trung
Thanh-Cong Le
University of Science
John von Neumann Institute
Vietnam National University
Ho Chi Minh city, Vietnam

E-Ro Nguyen
University of Science
Vietnam National University
Ho Chi Minh city, Vietnam

Mai-Khiem Tran
University of Science
John von Neumann Institute
Vietnam National University
Ho Chi Minh city, Vietnam

Tu-Khiem Le
Van-Tu Ninh
Cathal Gurrin
Dublin City University, Ireland

Minh-Triet Tran*
University of Science
John von Neumann Institute
Vietnam National University
Ho Chi Minh city, Vietnam

ABSTRACT

Building a retrieval system with lifelogging data is more complicated than with ordinary data due to the redundancies, blurriness, massive amount of data, various sources of information accompanying lifelogging data, and especially the ad-hoc nature of queries. The Lifelog Search Challenge (LSC) is a benchmarking challenge that encourages researchers and developers to push the boundaries in lifelog retrieval. For LSC'22, we develop **FIRST 3.0**, a novel and flexible system that leverages expressive cross-domain embeddings to enhance the searching process. Our system aims to adaptively capture the semantics of an image at different levels of detail. We also propose to augment our system with an external search engine to help our system with initial visual examples for unfamiliar concepts. Finally, we organize image data in hierarchical clusters based on their visual similarity and location to assist users in data exploration. Experiments show that our system is both fast and effective in handling various retrieval scenarios.

CCS CONCEPTS

• **Information systems** → **Search interfaces**; *Multimedia databases*;
• **Human-centered computing** → Interactive systems and tools.

KEYWORDS

Lifelog, Interactive Retrieval Systems, Semantic Embedding, Query Expansion

ACM Reference Format:
Nhat Hoang-Xuan, Hoang-Phuc Trang-Trung, Thanh-Cong Le, E-Ro Nguyen, Mai-Khiem Tran, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, Minh-Triet Tran. 2022. Flexible Interactive Retrieval SysTem 3.0 for Visual Lifelog Exploration at LSC 2022. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22), June 27–30, 2022, Newark, NJ, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3512729.3533013

1 INTRODUCTION

In 2021, we witnessed the rising popularity of video content on platforms such as TikTok, Instagram Reels, and YouTube Shorts. Along with the promise about the metaverse, they show that content is moving from simple images to more complex forms, such as video and virtual reality. A system that can handle these new forms of data, which is more costly computational and storage-wise, can undoubtedly provide great value. While the LSC'22 dataset[8] is not a video dataset on its own, it is similar to one, in terms of size and temporal meaning. It is greater than the previous edition (roughly 725,000 images compared to 183,299 [7]) and poses a significant challenge to system developers [8].

In recent years, Transformer [26] has become the prevalent architecture in both text and image domains. They have inspired the use of large collections of unlabeled data in training, which is easier to obtain. The approach is sometimes called self-supervised learning. When large image-text or video-text datasets become available, they give rise to vision-language pre-training. This approach creates large multi-purpose image-text models pre-trained on matching images to their captions instead of performing a specific task such as classification. These models are applicable to many downstream tasks such as Video-Text Retrieval [6], or even zero-shot Video Retrieval [16]. With the nature of being trained on image-text data, we believe that they are especially suitable for a cross-domain image-text retrieval task, which turns out to be exactly what LSC is.

Since the LSC competition centers around interactivity, this means that good systems also have practical value. In addition, they also hold a novice session to make sure the systems are easy to use, even for non-expert users. With that in mind, we seek to use the

*First author: hxnhat@selab.hcmus.edu.vn
Corresponding author: tmtriet@fit.hcmus.edu.vn

LSC '22, June 27–30, 2022, Newark, NJ, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9239-6/22/06…$15.00
https://doi.org/10.1145/3512729.3533013

large vision-language pre-trained models' representational strength and versatility to empower our search engine while at the same time simplifying user interaction through a better understanding of image sequence semantics.

Based on FIRST 2.0 [25], we revise and improve the functionalities and performance of our retrieval system, as well as integrate new components to FIRST 3.0. First, we enhance the semantic encoding for an image using CLIP [17]. We propose representing an image by a set of adaptive semantic embedding vectors, each corresponding to either the whole image or various regions of interest in different sizes. In this way, our system is expected to better capture the semantics of an image to search for concepts at varying levels of granularity. Second, we propose augmenting our system with an external search engine, such as Google, to find visual examples corresponding to unfamiliar concepts for our system to retrieve visually similar moments in the collection of images. Third, because of the vast amount of images, we utilize the clustering of images to shots, *i.e.* sequence of contiguous similar images, and scenes, *i.e.* similar shots in the same place at different time instants, to organize images in hierarchical clusters for efficient exploration.

The content of this paper is as follows. In Section 2, we briefly summarize existing approaches and systems for lifelog search challenge. Then, we present the overview and principal components of our retrieval system FIRST 3.0 in Section 3. Next, we illustrate several typical scenarios of usage for our system in Section 4. Finally, the conclusion and future work are discussed in Section 5.

2 RELATED WORK

Since the competition is in its fifth year, many systems have been proposed and improved over the years. Many of the systems also participate in other retrieval competitions such as Visual Browser Showdown [20] or ImageCLEF [4], each having a unique aspect in which they shine. Because our work focuses on the representation of image and text data, we also review other methods regarding that aspect.

In the LSC dataset, various tags (time, GPS location, visual concepts) are provided in the form of metadata. Most, if not all teams index these tags due to their availability and the intuitive method of querying based on tag. Methods such as [11] rely solely on this information, while [5] projects them onto a 3D cube, and [19] orders them into a graph and traverses it during the searching process. Some authors seek to generate additional tags using a pre-trained object detector, for example LifeSeeker [15] using ResNet101 [9], Myscéal 2.0 [22] utilizing DeepLabV3+ [3], and some ([10], [2]) used online API such as Microsoft Cognitive Service. While converting all information to tags makes them effective for indexing and searching, lifelogging data often contains errors that make it hard for task-specific detectors to work accurately.

Another common approach is to project the text and/or the image onto a common embedding space and use their distance as the relevance metric for retrieval. They can be used as local features in addition to tags in the case of Myscéal 2.0 [22] and LifeSeeker [15], or they can base their pipeline on it, such as Memento [1]. This idea is also adopted by SOMHunter [14] who uses the W2VV++ [12] model, combined with the recent CLIP [17]. [25] uses Faster R-CNN [18] as the image feature extractor, RoBERTa [13] as the text feature extractor, and they use custom layers to map them to a joint space. Since training a joint-embedding model is much more challenging than training a classifier and detector, large well-tested models are hard to find. However, recently such a model has become available [17]. We believe that many participants may switch to or incorporate CLIP [17] in their pipeline due to its strengths and ease of usage.

The two mentioned approaches often complement each other, so we make use of both: we use tags when we are confident about their correctness (e.g., time and location) so we can quickly narrow down the search space, and we leverage the joint embedding to navigate through the remaining candidates with query expansion techniques that we describe later.

3 METHOD

Our system is designed with simplicity in mind. We leverage the strong representation learning capabilities of recent large models such as CLIP[17] to reduce the complexity of user interactions required. Furthermore, we make use of the common embedding space to find similar examples to our queries to speed up the searching process. This is similar to "exploring" the embedding space, and through this process we also have an idea of the quality of the space that the model has learned.

3.1 System Design Principles

Being a multi-module system, the design and inclusion of components in FIRST are influenced by the following principles:

- **Flexibility**: As the F in FIRST, we place flexibility at a high priority in our design. With the goal of becoming a general system that can support various user needs, we develop our system in a manner such that the componenets are not tightly coupled and can be easily swapped in-and-out in cases where an upgrade or replacement is desired. This structure also allows us to quickly evaluate and test the various components to find the best combination for a particular use case.
- **Scalability**: As the size of the dataset will only grow over years, for a system to remain relevant, it has to be able to scale. There are three aspects of scaling that we consider in our system: *storage efficiency*, *retrieval performance*, and *ease of browsing*. It is easy to see that these factors can influence each other, for example, effective retrieval might require additional information saved along with each image, which reduces storage efficiency. To achieve this, we use scalable databases as the backbone of our retrieval system, in addition to data structures that implement hierarchical indexes for browsing.
- **Openness**: As much as we want it to be, our system can never cover all existing concepts. However, by allowing the retrieval process to be guided by external knowledge, we can have an effectively unlimited pool of understood concepts, while leveraging the effectiveness of our searching tool. We obtain this by modelling similarity and dissimilarity between images/texts; this feature enables the possibility of using of an external example as a prototype, then performing query expansion based on it. We also seek to advance further from

simply defining concepts to be objects that appear in an image, instead we look to model higher semantic meaning such as events (e.g., birthday party), emotions, activities, and more. We present more examples of this feature later on in the paper.

3.2 System overview

Figure 1 demonstrates the overview architecture of our retrieval system FIRST 3.0. The current system is developed from FIRST 2.0 [25]. The top and bottom layers are the platforms that allow the flexible integration of different modules for query processing, query expansion, visualization and interaction [24, 25].

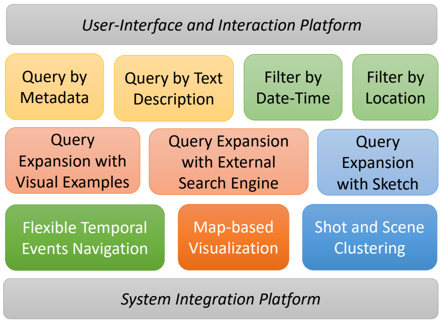

User-Interface and Interaction Platform
Query by Metadata
Query by Text Description
Filter by Date-Time
Filter by Location
Query Expansion with Visual Examples
Query Expansion with External Search Engine
Query Expansion with Sketch
Flexible Temporal Events Navigation
Map-based Visualization
Shot and Scene Clustering
System Integration Platform

Figure 1: Overview of retrieval system.

Our system supports users to filter moments by date-time and location, and retrieve images by metadata and text description. For a text description query, we propose the mechanism for attention-based embedding enrichment (see Section 3.4) to capture interesting semantic features of an image in different image regions at various levels of granularity. We also support nested queries and the combination of different queries. We provide several query expansion modules to further assist users. The query expansion with visual examples [25] allows users to search for visually similar moments from a given image. We propose a new idea for query expansion with the assistance of external search engine to find unknown/unfamiliar concepts (see Section 3.5). We also provide a simple sketch-based retrieval [23] so that users can quickly sketch out the scene of interest.

Besides the regular retrieval interfaces, we revise and enhance three modules for user interface and interaction, including a flexible temporal events navigation, map-based visualization, and shot/scene clustering.

3.3 Pre-processing and indexing

As the lifelogging data is a large collection of images with low information density, we seek to efficiently reduce its size prior to constructing an index. We achieve this by applying a few pre-processing steps, as illustrated in Figure 2:

- **Filtering**: We filter out images that are blurry due to motion or obstruction, as they do not contain useful information.
- **Normalization**: Sometimes the images can be rotated by multiples of 90 degrees, so we normalize them to ensure our next processing steps work correctly.

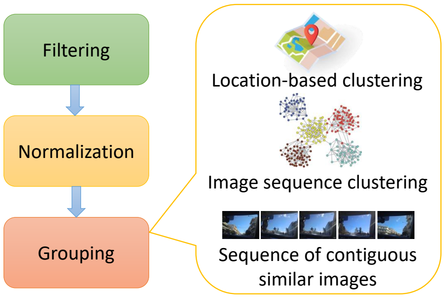

Filtering
Normalization
Grouping
Location-based clustering
Image sequence clustering
Sequence of contiguous similar images

Figure 2: Pre-proceessing and clustering.

- **Grouping**: We group segments of very similar images caused by stationary viewpoint (which occurs very frequently) into one so they do not overcrowd the search results.

The pre-processing steps are implemented using both embedding-based comparison and simple computer vision techniques, yet they greatly reduce the number of images left and make sure they are meaningful, converting raw data to information-packed data.

In Figure 2, we also demonstrate the main steps in grouping images into more meaning groups. For grouping images, we first gather *contiguous* images that are visually similar into a shot, *i.e.* a sequence of similar images. For each shot, we select a key frame, usually the first image in that sequence, to represent that shot and assign the GPS information to that shot from its images. We can also select multiple key frames if we enlarge the shot with several contiguous similar sub-shots. Then we cluster key frames of all shots to link shots taken at the same scene in different time instants. Finally, thanks to the GPS information of images and shots, we can group images in shots based on their locations. We also exploit the hierarchical relationships of locations and places, such as buildings, cities, or countries, to enhance the functionality of our retrieval solution.

In LSC'22, each image is also accompanied with metadata such as time and location, texts present in the image, and visual concepts. As with previous years, these information are vital to finding the required shots, therefore we indexed them in our database. We also enrich the concepts associated with each image using the CLIP model[17] and Conceptual Captions [21] tags.

3.4 Attention-based embedding enrichment

A traditional approach for lifelog retrieval is to extract concepts from an image so that it can be indexed. The extracted concepts can be entities appearing in the image, type of place, type of action, etc. However, this approach depends on the concept detectors for known concepts in a pre-defined dictionary . Therefore, this method is not appropriate to search for new concepts that are not available in that dictionary.

Keep in mind the openness for our retrieval system, we aim to represent an image with feature vectors that can be used to match with new concepts. For each image in the dataset, we extract a high-dimensional representation using CLIP [17]. This embedding is a good general descriptor of the image, and it is close to the main features (concepts) of the image. In this way, our system can

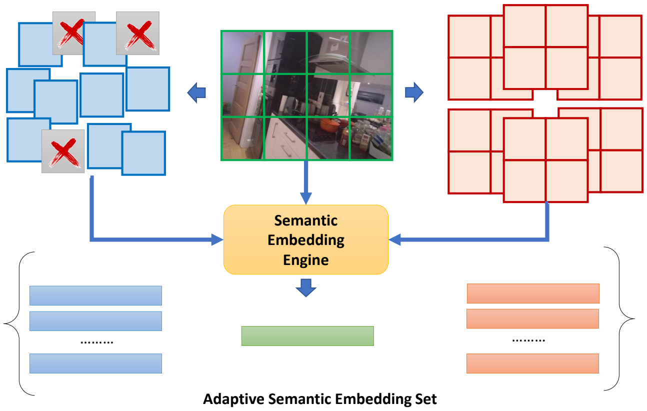


Figure 3: Image decomposition and Adaptive semantic embedding set.

support users search for simple concepts related to entities in an image (such as *chair*, *TV*, *sandwich*, *etc*) as well as more abstraction concepts (such as *a lecture class*, *a wedding ceremony*, *etc*).

However, the lifelogging data contains very similar scenes, in which a subtle change in the content of the image can differentiate one image from hundreds of similar ones (e.g., the content of the TV at that time). Therefore, apart from the general concepts in the image, we are also interested in more subtle and local features.

Furthermore, since the original size of the images is 1024 × 768 and the CLIP model encodes images of size 224 × 224 or 336 × 336 depending on the model used, they have to be down-sampled in order to be encoded by CLIP, which can cause loss of information. Furthermore, it would not be sufficient to query, *i.e.* to match, an object or concept appearing in a small portion of an image from the global embedding vector of that image. Hence, it should be necessary to encode various important regions of an image at different levels of details to assist retrieving different concepts at various levels of granularity in an image.

Due to the above reasons, we seek to add additional information besides the overall image's embedding. To achieve this, we select smaller crops of the image and encode them along with the original image. This is similar to the attention mechanic in the sense that we focus on "interesting" regions in the image. The regions to be selected are usually the ones that contain salient objects that can define the scene. In this way, we can represent an image with an adaptive semantic embedding set.

Figure3 demonstrates our idea to decompose an image into multiple patches, corresponding to different levels of detail, and generate an adaptive semantic embedding set corresponding to that image. In our implementation for FIRST 3.0, for simplicity, we represent an image (with the aspect ratio of 4:3) as a grid with 4 × 3 square cells. Then we construct multiple patches with the size of 1 × 1 and 2 × 2 cells, which can be overlapped. Finally, we remove unimportant patches and encode information-rich patches into semantic embedding vectors. The adaptive semantic embedding set is the collection of semantic embedding vectors of the full-size image and its exciting patches.

3.5 Visually-similar image searching

As stated in Section 3.1 in the principle of *Openness*, we use similarity modelling to extend the capabilities of our system. This is a feature that many other methods also seek due to its usefulness in the retrieval setting [15] [14]. In our system, we leverage the strong representational capability of CLIP [17], which is demonstrated to be effective even in the zero-shot setting [16]. Because CLIP was trained on image-caption classification task which is a form of contrastive training, we believe it has learned to differentiate between images based on the concepts existing in the images, and so it can encode the image regardless of the content. More importantly, this allows us to simply define the distance between two images as the cosine distance between their embeddings. With this definition, we can quickly find an image that is similar to a given image in a database. This gives rise to the ability to search using visual examples.

In multiple cases, there are some concepts cannot be well described using words, are unknown for the user, or are unavailable in the training corpus for the text encoder. Such instances are present in previous editions of the LSC. While trying to expand the pre-defined dictionary helps in all cases, it requires much effort to identify the needed concepts and also comes with storage and computing costs. We deviate from this paradigm by two means: utilizing *external* systems and modelling *deep* image semantic.

With the ubiquitous amount of data available on the Internet, we believe that any possible concept, possibly along with an image-text correspondence, exists and can be found with an appropriate tool, such as Google Search. We can query those tools to find an example of such concept, and use it as a starting point for our retrieval process. With our visual comparison capabilities mentioned earlier, we support the use of an external image as a prototype for query, as demonstrated in Figure 4. We can look for *coffee machine* with visual examples suggested by an external search engine. This approach allows us to broaden the scope of searching beyond our existing concepts, and utilizing the strengths of other systems, while also producing a natural and intuitive searching process. This feature works well with our adaptive embedding described in Section 3.4, as the particular unknown/unfamiliar concept usually only occupies a portion, or even a tiny bit of an image, and therefore focusing on it greatly helps with "matching" it to the available prototype.

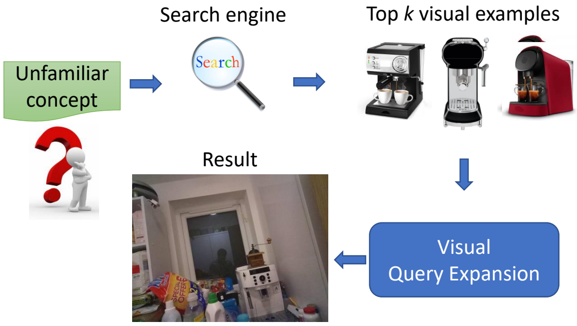
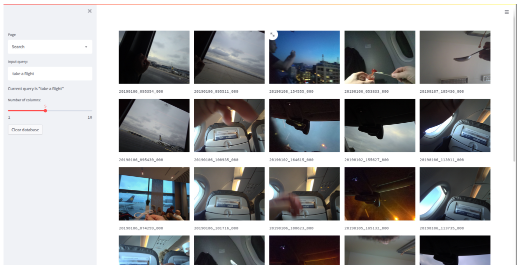


Figure 4: Retrieval of an unfamiliar concept with the assistance of visual search engine.

Many works [22], [25] enrich images with metadata tags from pre-trained state-of-the-art object detectors. While this is intuitive and effective and we do this ourselves too, we propose a further advancement by modelling more abstract concepts from the image, such as events. This can be achieved through a number of ways, one of them is breaking down or associate an abstract concept with simple concepts or objects that can be searched for. An example of this is instead of searching for "teaching" which is difficult to define, we can try to look for whiteboards, people in a room, desks, etc. As another avenue, we note that recent advances in representation learning have make this more possible than ever; we show that since CLIP was trained on image-caption pairs, it has some (limited) ability to understand a scene in the same way a human does, and we can directly leverage this to empower our searches. By leveraging general representations instead of task-specific representations, advances to these models also empower our search engine without intervention. We believe this direction may prove useful and make future search engines simpler yet more powerful.

3.6 System interface and navigation



Page

Search

Input query:

take a flight

Current query is "take a flight"

Number of columns: 5

Clear database

Figure 5: The interface of our system.

The interface of our system can be seen in Figure 5. In summary, our system is based on FIRST 2.0 [25], with the following enhancements:

- **Better scene clustering**. Despite having implemented scene clustering, our previous version suffers from repetitive scenes. In our new version, we improve this by using CLIP features and using heuristics to merge scenes. The grouping can be collapsed/expanded as needed to make sure browsing is fast by default, while additional details can still be obtained in case of a need for high recall. We also support other strategies for clustering such as geographical proximity.
- **Flexible temporal navigation**. We support quickly browsing through the photos of a day with an adjustable level of details, and with an enhanced user interface. We also support browsing through moments before and after a moment in a specific results, so that users can check for related activities or places of a query.
- **Local/prototype visual search**. With the ability to search using external examples and the ability to focus on local regions described in Sections 3.4 and 3.5 respectively, visual-based searches have become more useful than ever. The idea of googling an unknown concept is also logical to a novice user. The search bar accepts URL of images, so the process is quick and convenient.
- **Better representation**. With the adoption of CLIP instead of ResNet [9] and the shift in paradigm mentioned in Section 3.5, the representation quality has increased significantly and search quality has improved, as well as enabling a host of other features. Since we use both image encoder and text encoder from CLIP, the joint embedding is more aligned and easier to work with.

4 CASE STUDY

In this section, we describe some specific usage scenario to demonstrate how to best use our system, as well as its strengths.

Scenario 1: *I was looking at a lead soldier in a mall, next to some clothes.*

We think of two strategies to approach this query, either directly searching for "*lead soldier*", or imagine a typical lead soldier and try to describe it. For the second strategy, we might attempt to look for a standing man, wearing red shirt and a top hat.

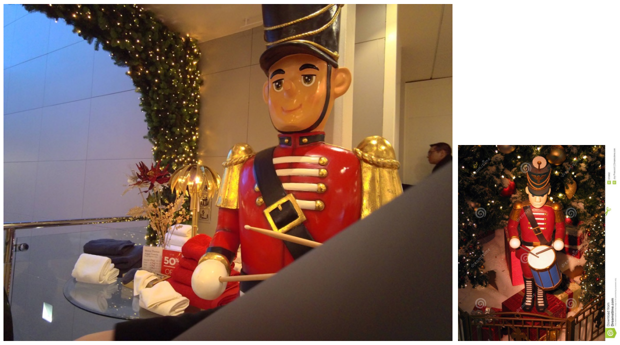


Figure 6: The first scenario. The target image is on the left, while the image on the right is the prototype we took from Pinterest.

The results of both approaches are shown below.

a lead soldier standing ✗
a lead soldier wearing red shirt ✗
a lead soldier wearing red suit ✓
a man wearing red suit ✗

We observe that searching for "man" or "red suit" yields too much results and not the one we are looking for, and even in the query that successfully found the target image, the results are inconsistent. Instead, we can search for a concrete example on the Internet and use it directly for our search, as shown in Figure 6. This gives the target image as the top-2, and at the same time yield 2 other images with a lead soldier in it in the top-15, which none of the shown queries were able to. As mentioned above, we only need the URL of the image, so the process itself is fairly quick, and it is much faster than the try-and-error approach of query engineering.

Scenario 2: *I was taking a photo of a man sitting at a table. I took a flight to this city 3 days ago. I was in Greece then.*

In this example, we would like to illustrate our system's capability to retrieve a moment with the activity or story in an image, instead of looking for only entities/objects appearing in it. With

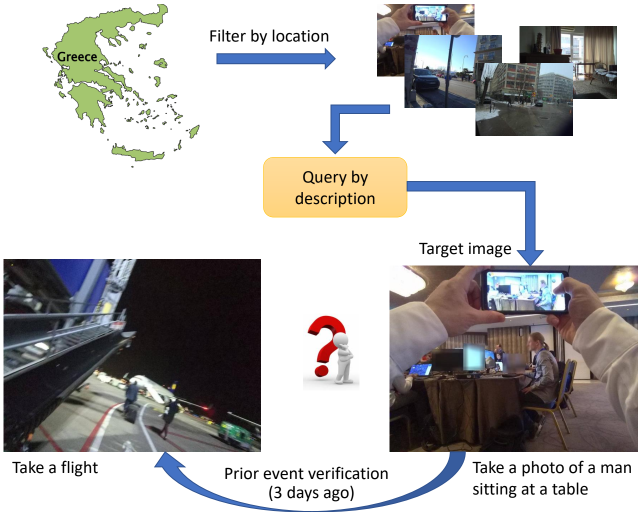

Greece
Filter by location
Query by description
Target image
Take a flight
Prior event verification (3 days ago)
Take a photo of a man sitting at a table

Figure 7: The second scenario. The target image can be identified from description, location filtering, and prior event verification.

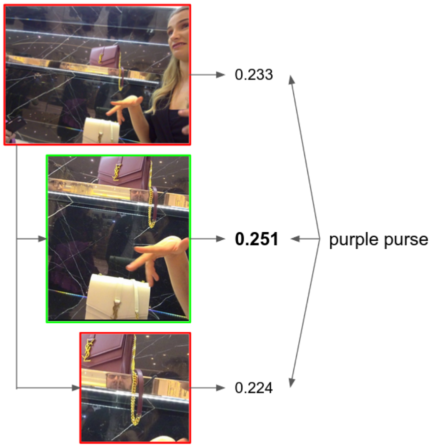


Figure 8: The third scenario. The target image is shown at the top, following by a 2×2 patch and a 1×1 patch containing the purse. The images are not scaled equally.

the first sentence in the query description, rather than looking for the moment when we can see a smartphone or camera in a photo, we can simply search with the text description "*I was taking a photo of a man sitting at a table*".

If there are multiple moments, we can verify the event of interest using the second sentence in the query. Three days can be a long duration and it would be inconvenient and inefficient for a user to navigate through a long sequence of images just to verify the previous event of taking flight and arriving at a new city. Our system assists users with a flexible navigation mechanism and grouping similar images into shots, thus the system helps users quickly confirm the moment happening 3 days ago.

Finally, if we wait until the last minute to exploit the last sentence in the query description, we can narrow down the moments occurring in Greece. We can simply set the filter on the location ("*in Greece*") and query with the description from the first sentence, and we can successfully identify the target moment, as illustrated in Figure 7.

Scenario 3: *I was in a shop, talking to a woman in front of some purses. One of them was purple and the other one was white.*

For this query, we could try to locate the shop through geolocating or finding other shopping moments and browse for a fashion shop that sells purses. We believe that there are few images of a purple purse, so we focus on this information. However, from the description, we know that the purse does not occupy the whole image, this can make it difficult to find it if we only encode the whole image.

Indeed, the local features mentioned in section 3.4 help us in this case, as the correct crop can better highlight the purse. Note that the varying sizes matter, as the 1×1 patches fail to capture the purse, and only the 2×2 patch is able to fully capture it. Being able to centralize on the purse while ignoring other objects pushes it closer to the "purse" concept and allows it to be found. The scenario is depicted in Figure 8. The scores represent in the figure are the similarity (higher is better) to the query phrase "*purple purse*". In this case, only the 2×2 patch is close enough to make it to the top results, the other images are too far down the ranklist.

5 CONCLUSION

In this paper, we present FIRST 3.0, a newly rebuilt version of our previous systems [24, 25]. Our intention is to ensure that our system can handle large data while retaining its flexible architecture. Furthermore, we propose to use CLIP [17] in a novel way so that our system can adaptively capture semantics at different levels of detail in an image. Each image is associated with a set of semantic embedding vectors to represent the image at various levels of granularity, from the whole picture to patch sizes. To deal with unfamiliar concepts, we propose an augmentation for our system by using an external search engine to find initial visual examples corresponding to a new concept. Furthermore, the user interface of our system allows interaction with an adjustable level of granularity.

To assess whether users can use our tool easily and logically, we need to conduct further novice and expert evaluation. To achieve our goals, our system also needs further realization and revisions, which can take advantage of the flexibility architecture. We hope that our ideas lead to better search engines, both in terms of performance and accessibility.

ACKNOWLEDGMENTS

This 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] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2021. Memento: A Prototype Lifelog Search Engine for LSC'21.. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[2] Wei-Hong Ang, An-Zi Yen, Tai-Te Chu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. An Interactive Approach for Multimodal Lifelog Retrieval through Concept

Recommendation. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11211. Springer, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49

[4] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Matthias Lux, Minh-Triet Tran, Tu-Khiem Le, Van-Tu Ninh, and Gurrin Cathal. 2019. Overview of ImageCLEFlifelog 2019 : Solve My Life Puzzle and Lifelog Moment Retrieval. *Working Notes of {CLEF} 2019 - Conference and Labs of the Evaluation Forum,Lugano, Switzerland. 09 Sept 2019* (2019), 9–12.

[5] Aaron Duane and Björn Þór Jónsson. 2021. ViRMA: Virtual Reality Multimedia Analytics at LSC 2021. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[6] Zijian Gao, Jingyu Liu, Sheng Chen, Dedan Chang, Hao Zhang, and Jinwei Yuan. 2021. CLIP2TV: An Empirical Study on Transformer-based Methods for Video-Text Retrieval. *CoRR* abs/2111.05610 (2021). arXiv:2111.05610 https://arxiv.org/abs/2111.05610

[7] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC’21. In *ICMR ’21: International Conference on Multimedia Retrieval, Taipei, Taiwan, August 21-24, 2021*. ACM, 690–691. https://doi.org/10.1145/3460426.3470945

[8] 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 *ICMR ’22: International Conference on Multimedia Retrieval, Newark, NJ, USA, June 27-30, 2022*. ACM. https://doi.org/10.1145/3512527.3531439

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[10] Silvan Heller, Ralph Gasser, Sanja Popovic, Luca Rossetto, Loris Sauter, Florian Spiess, Heiko Schuldt, and Mahnaz Parian. 2021. Interactive Multimodal Lifelog Retrieval with vitrivr at LSC 2021. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[11] Andreas Leibetseder and Klaus Schoeffmann. 2021. lifeXplore at the Lifelog Search Challenge 2021. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[12] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2VV++: Fully Deep Learning for Ad-Hoc Video Search. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM ’19)*. Association for Computing Machinery, New York, NY, USA, 1786–1794. https://doi.org/10.1145/3343031.3350906

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[14] Jakub Lokoč, František Mejzlík, Patrik Veselý, Miroslav Kratochvíl, and Tomáš Souček. 2021. Enhanced SOMHunter for Known-item Search in Lifelog Data. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[15] Thao-Nhu Nguyen, Van-Tu Ninh, Tu-Khiem Le, Cathal Gurrin, Thanh-Binh Nguyen, Minh-Triet Tran, Annalina Caputo, and Graham Healy. 2021. LifeSeeker 3.0 : An Interactive Lifelog Search Engine for LSC’21. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[16] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. 2021. A Straightforward Framework For Video Retrieval Using CLIP. *CoRR* abs/2102.12443 (2021). arXiv:2102.12443 https://arxiv.org/abs/2102.12443

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research)*, Marina Meila and Tong Zhang (Eds.), Vol. 139. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[18] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91–99. https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[19] Luca Rossetto, Matthias Baumgartner, Ralph Gasser, Lucien Heitz, Ruijie Wang, and Abraham Bernstein. 2021. Exploring Graph-querying approaches in LifeGraph. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[20] Klaus Schoeffmann, David Ahlström, Werner Bailer, Claudiu Cobârzan, Frank Hopfgartner, Kevin McGuinness, Cathal Gurrin, Christian Frisson, Duy-Dinh Le, Manfred Del Fabro, Hongliang Bai, and Wolfgang Weiss. 2014. The Video Browser Showdown: A Live Evaluation of Interactive Video Search Tools. *International Journal of Multimedia Information Retrieval (MMIR)* 3 (06 2014), 113–127. https://doi.org/10.1007/s13735-013-0050-8

[21] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[22] Ly-Duyen Tran, Manh-Duy Nguyen, Hyowon Lee, Thanh Binh Nguyen, and Cathal Gurrin. 2021. Myscéal 2.0: A Revised Experimental Interactive Lifelog Retrieval System for LSC’21. In *Proceedings of the 2021 ACM Workshop on the Lifelog Search Challenge, LSC21*. Taipei, Thailand.

[23] Minh-Triet Tran, Nhat Hoang-Xuan, Hoang-Phuc Trang-Trung, Thanh-Cong Le, Mai-Khiem Tran, Minh-Quan Le, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. 2022. V-FIRST: A Flexible Interactive Retrieval System for Video at VBS 2022. In *MultiMedia Modeling - 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6-10, 2022, Proceedings, Part II (Lecture Notes in Computer Science)*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Huynh Thi Thanh Binh, and Benoit Huet (Eds.), Vol. 13142. Springer, 562–568. https://doi.org/10.1007/978-3-030-98355-0_55

[24] Minh-Triet Tran, Thanh-An Nguyen, Quoc-Cuong Tran, Mai-Khiem Tran, Khanh Nguyen, Van-Tu Ninh, Tu-Khiem Le, Hoang-Phuc Trang-Trung, Hoang-Anh Le, Hai-Dang Nguyen, Trong-Le Do, Viet-Khoa Vo-Ho, and Cathal Gurrin. 2020. FIRST - Flexible Interactive Retrieval SysTem for Visual Lifelog Exploration at LSC 2020. In *Proceedings of the Third ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2020, Dublin, Ireland, June 8-11, 2020*, Cathal Gurrin, Klaus Schöffmann, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, and Wolfgang Hürst (Eds.). ACM, 67–72. https://doi.org/10.1145/3379172.3391726

[25] Hoang-Phuc Trang-Trung, Thanh-Cong Le, Mai-Khiem Tran, Van-Tu Ninh, Tu-Khiem Le, Cathal Gurrin, and Minh-Triet Tran. 2021. Flexible Interactive Retrieval SysTem 2.0 for Visual Lifelog Exploration at LSC 2021. In *Proceedings of the 4th Annual on Lifelog Search Challenge, LSC@ICMR 2021, Taipei, Taiwan, 21 August 2021*, Cathal Gurrin, Klaus Schoeffmann, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoc, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy (Eds.). ACM, 81–87. https://doi.org/10.1145/3463948.3469072

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.