NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001335

MEMORIA: A Memory Enhancement and MOment RetrIeval Application at the NTCIR-17 Lifelog-5 Task

Ricardo Ribeiro
IEETA/DETI, Universidade de Aveiro
Portugal
rfribeiro@ua.pt

Alexandre F. P. Gago
IEETA/DETI, Universidade de Aveiro
Portugal
alexandre.gago@ua.pt

Bernardo F. M. Kaluza
IEETA/DETI, Universidade de Aveiro
Portugal
bernardokaluza@ua.pt

Josefa N. S. Pandeirada
WJCR, Universidade de Aveiro
Portugal
josefa@ua.pt

António J. R. Neves
IEETA/DETI, Universidade de Aveiro
Portugal
an@ua.pt

ABSTRACT

In recent years, the practice of continuously recording and collecting information about several aspects of individuals' lives has gained increased popularity. This practice, known as lifelogging, serves multiple purposes, including personal health monitoring and enhancement as well as recording day-to-day activities in hopes of preserving some memories. An essential aspect of this practice lies in the gathering and analysis of image data, offering valuable insights into an individual's lifestyle, dietary patterns, and physical activities. The NTCIR Lifelog Challenge presents a unique opportunity to delve into the latest advancements in lifelogging research, particularly in the field of image retrieval and analysis. Researchers are encouraged to present their methodologies and participate in lifelog retrieval challenges. Consequently, these challenges allow research teams to assess the efficiency and accuracy of their developed systems using a multimodal dataset derived from an active lifelogger's 18 months of continuous lifelogging data. This paper presents the current version of MEMORIA, a computational tool that provides an intuitive user interface with several options that allow the user to upload images, explore the segmented events, and perform image retrieval, namely images for the NTCIR Lifelog event. This version of MEMORIA incorporates natural language search capabilities for information retrieval, offering options to filter results based on keywords and time periods. The system integrates image analysis algorithms to process visual lifelogs. These algorithms range from pre-processing algorithms to feature extraction methods, to enrich the annotation of the lifelogs. The paper also includes experimental results of the image annotation methods used in MEMORIA, as well as some examples of user interaction.

KEYWORDS

lifelog, lifelogging, image processing, image annotation, data retrieval, object detection, Machine Learning, Information Systems

TEAM NAME

MEMORIA - A Memory Enhancement and MOment RetrIeval Application

SUBTASKS

Lifelog Semantic Access SubTask - LSAT

1 INTRODUCTION

The rapid advancements in technology in recent years have given rise to the use of devices, such as smartphones, smartwatches, and other wearable technologies. These devices enable the gathering of data about our daily activities and behaviors. This data acquisition process, often called lifelogging, can be carried out actively or passively, resulting in the accumulation of digital records known as lifelogs [14]. These records can be used in several ways to gain insights into an individual's health, memories, behaviors, and daily routines. However, developing a system capable of efficiently handling these data can be very challenging due to the high amount of data, coupled with the diversity of devices used to acquire data, that make these records highly heterogeneous. Additionally, the data can take various forms, including images, audio, coordinates, biometric data, and even documents.

The effective use of these records requires efficient organization, processing, and data retrieval, a challenging and demanding task given its multimodal nature. Usually, responsibility for these tasks falls into a lifelogging system, which extracts valuable insights from each record to facilitate the search and retrieval of specific events within the digital memory archive of the lifelogger.

The management and retrieval of accumulated lifelogs pose a multifaceted challenge that has gained increased attention in recent years. To encourage research to challenge and establish benchmarks for assessing developed systems, various events have been introducing lifelogging retrieval tasks. Among these, the NTCIR workshop proposes a lifelog retrieval task [24] that aims to advance state-of-the-art research in lifelogging as an information retrieval application. The NTCIR workshop proposes the Lifelog Semantic Access Task (LSAT) is a known-item search task where participants have to retrieve several specific moments in a lifelogger's life. These moments are defined as semantic events or activities that happen throughout the day. Using the dataset provided by the organizers, each team will compete against each other to demonstrate which lifelogging system has the best speed and accuracy and be crown the winner.

In this paper, we present the MEMORIA (Memory Enhancement and MOment RetrIeval Application) for the participation in the LSAT task of NTCIR-17 lifelog [24]. The MEMORIA system is a computational web-based tool providing an intuitive user interface with several options that allow the user to upload, explore,

and visualize lifelogs, and perform image retrieval, among other functionalities [13].

Following this introduction, Section 2 provides an overview of related work that fostered research in lifelogging. Section 3 presents the MEMORIA system overview, including the technologies that it employs to extract annotations and captions from images. Section 4 presents and discusses some of the system's results. Finally, in Section 5, a conclusion and ongoing work are detailed and explored.

2 RELATED WORK

In recent years, numerous lifelog applications have emerged in academic literature. Many of them actively engage in competitions designed to evaluate the latest advancements in the field of lifelog retrieval. Among these competitions, the Lifelog Search Challenge (LSC) stands out as being an annual competition that is dedicated to promoting research in personal lifelog retrieval. Each participating team, similarly to the NTCIR lifelog task [24] competes against each other in lifelog retrieval tasks. In LSC23's event [6], several systems were presented.

Memento 3.0 [1] employs image-text embeddings derived from two distinct CLIP[12] models to create a ranking mechanism that aims to merge the similarity scores between an image and the query by aggregating the ratings produced by multiple models' output.

Voxento 4.0 [2] offers a voice-based retrieval mechanism, which enhances the accessibility of lifelog retrieval. It incorporates data preprocessing and cleaning techniques for the dataset and leverages the CLIP model to extract image features. The system also provides the option of using a text-based search to complement the voice-based method. By using textual information found in images the system can also detect the environment of the image.

LifeXplore [15] offers a system that relies on the extraction of visual concepts from images. It uses a CLIP [12] model to generate text embeddings and image similarity, CRAFT [4] for text recognition within images, YOLOv7 [20] to detect objects, and an EfficientNet B2 to generate semantic concepts. All these technologies work together to generate tags that help in the retrieval process of images.

MemoriEase [19] combines concept-based and embedding-based retrieval approaches. The system uses BLIP [7] for its embedding-based retrieval method, which aims to minimize the semantic gap between images and text queries. In addition, the concept-based retrieval approach utilizes full-text search within Elasticsearch to retrieve images that share visual concepts similar to the keywords in the query.

MyEachtra [16], an improved system prom MyScéal [17], shifts the focus from images to events as retrieval units. These events are defined by the utilization of location metadata, as well as visual and temporal disparities between consecutive images. Given a query, the system shows the top 3 events that proved to be the most relevant to the text query. This system uses CLIP-H/14 to generate the image embeddings as well as an additional source of location metadata, VAISL [18] to help in the event segregation.

These systems provide valuable insights into useful technologies on lifelog retrieval systems while highlighting the prevailing trends and challenges in this field. Each system presents different approaches that provide insightful views on how to tackle this complex problem.

3 MEMORIA OVERVIEW

Figure 1 shows a general representation of the latest MEMORIA system version developed. It was designed and developed as a web application capable of storing, organizing, analyzing, visualizing, and retrieving personal lifelog data, allowing users to create individual accounts in order to preserve their privacy. The architecture and structure of MEMORIA have been improved over time and were designed with distinct modules to optimize the lifelogging experience. Each module serves a unique function, contributing to the effective management of lifelog data; these modules include storage, organization, annotation, retrieval, and visualization. The MEMORIA system also incorporates other functionalities to visualize and analyze lifelogs, providing users with intuitive tools to interact with their lifelog data. These features enhance the users' ability to understand and interpret their lifelog content.

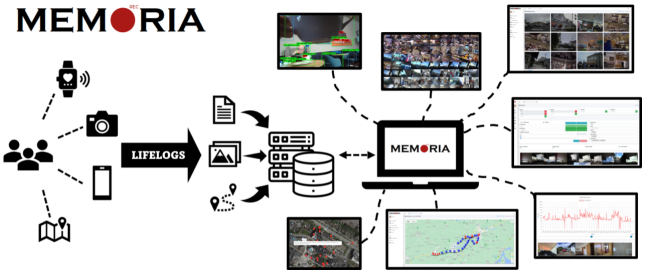


Figure 1: General representation of the MEMORIA lifelogging system.

3.1 Semantic Location Annotation

As the dataset used for the NTCIR-17 Lifelog-5 Task [24] is the same as the one used in the LSC'23 challenge [6], it incorporates essential metadata such as time, and GPS coordinates, among others. MEMORIA can filter and cluster GPS coordinates daily as exemplified in Figure 2. These location clusters are formed using the HDBSCAN clustering algorithm [9], and subsequently, a geocoding reverse API is used to convert the clustered locations to addresses that enrich annotations.

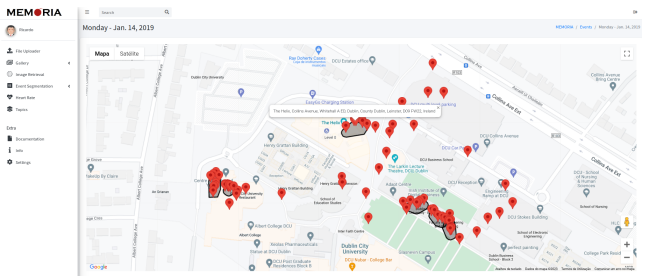


Figure 2: Example of clusters annotated by MEMORIA.

Table 1: Different annotation algorithms integrated into MEMORIA.

Computer Vision Task	Extracted Annotations
Object Detection	Classes and location of objects
Object Understanding	Descriptive sentences and location of objects
Optical Character Recognition	Recognition of optical characters
Scene Understanding	Categorization of the setting where the scene takes place, as well as its attributes
Automatic Caption Generation	Natural language description of the contents of the image

MEMORIA leverages a deep learning model trained on the Geolife GPS Trajectory dataset [22] to discern five distinct modes of transportation: walking, biking, driving, bus, and train. Furthermore, MEMORIA uses a temporal synchronization mechanism to annotate location and transport mode information within the dataset images to associate the GPS metadata with the images. This system allows searching for images based on transportation mode and location which is essential for the Lifelog Semantic Access SubTask - LSAT.

3.2 Image Annotation

Automatic image annotation refers to the automated task of assigning labels and additional information to digital images. Its primary objective is to allow for images to be retrieved similarly to how we search for text. Nonetheless, achieving this goal requires the extraction of meaningful information from a collection of lifelog images which can help to close the gap between images and text-based queries. This can be a difficult task as an image contains various types of information that can be harnessed to create semantic meaning.

To achieve this goal, MEMORIA offers several types of image annotation, as seen in Table 1 and Figure 3: Object detection, Object Understanding, Optical Character Recognition, Scene Recognition, and Caption generation.

3.2.1 Object Detection and Understanding. The object detection task focuses on identifying specific objects within images or videos. Typically, when searching for an image, it is common to search for a specific object, making it crucial to associate each image with the objects it contains. The current version of MEMORIA uses the YOLOv7[20] object detection model which offers fast and accurate object detection and returns single-word labels for recognized objects.

Similarly to the previous task, object understanding is centered around identifying objects within the lifelogs. However, unlike the previous task, this one is responsible for generating short textual descriptions of objects. If, for example, an object detection model such as YOLOv7 detects "chair" and "person", the object understanding model detects "person sitting on a chair". However, while object understanding models provide more comprehensive information, they sometimes can miss some relevant objects. Furthermore, they are slower and less efficient than the object detection models, which makes the use of object detection models still needed especially for more real-time applications. The model used for object understanding is GRiT[21], a powerful tool for generating detailed descriptions in free-form text about objects. What sets this model apart from the rest is its ability to be trained with a wide range of free-form text styles, such as class names and descriptive sentences that encompass object quantities, attributes, and actions. Its versatility and remarkable accuracy make it a great solution for the generation of annotations.

MEMORIA combines GRiT and YOLOv7 annotations and uses an algorithm based on overlapping areas to reduce redundancy while prioritizing GRiT annotations. The remaining annotations are used by MEMORIA. This process of redundancy reduction can be seen in Figure 4.

3.2.2 Optical Character Recognition. Optical character recognition (OCR) is another important tool in generating image annotations. In our daily lives, we often encounter objects with text written on them like labels, signs, etc. As such, it is understandably common that many images will include text. Extracting this text for the image annotations would be extremely useful in the image search as it can provide contextual clues about the scene, location, time, or actions depicted. CRAFT [4] and a text recognition model [3] were used to extract text from an image. CRAFT is a model that extracts portions of the image that have text and the text recognition model takes the cropped images and infers the characters that are written in it.

One of the main strengths of CRAFT is its ability to accurately detect and locate curved or deformed regions of text in images. This capability is critical as, often times, text in the lifelog images has blurred, small, or deformed text that could be useful to create some image annotations.

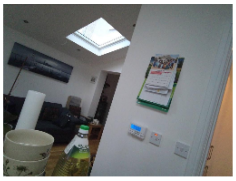
3.2.3 Scene Recognition. Scene recognition plays an important part when retrieving moments and is a very challenging problem as an image can present various objects and layouts which can lead to ambiguity when trying to classify the image. MEMORIA uses a pre-trained model provided by Zhou et al. [23] trained on the Places365-Standard dataset for this task which identifies scenes such as “restaurant” and “enclosed area”.

3.2.4 Caption generation. To further enrich the image annotations, MEMORIA uses ClipCap [11], a model that generates unique captions for each image. It uses the multimodal embedded model, VIT-B/32 CLIP [12], trained with the COCO [8] dataset, to generate image embeddings that are then used to generate a caption for the given image. The annotations from ClipCap differ from those of the other models because it does not generate annotations for objects but rather for the image as a whole, such as “A man sitting at a table with a coffee mug in front of him”.

3.3 Segmentation of events

Human beings can remember moments of their lives as episodes or events and doing this allows us to recall life experiences in the form of these moments. An event can be described as a period of time delimited by a starting and finishing point. Usually, lifelog data is captured passively and continuously, and organizing this data into temporal segments each with its different events is commonly referred to as event segmentation [5, 13]. MEMORIA uses

Types of annotations	MEMORIA
Object detection	[(car, 0.621), (truck, 0.663)]
OCR	[]
Caption generation	"view from the driver's seat."
Deep annotation	[("red truck in rear", 0.554), ("the rear view mirror on the car", 0.554)]

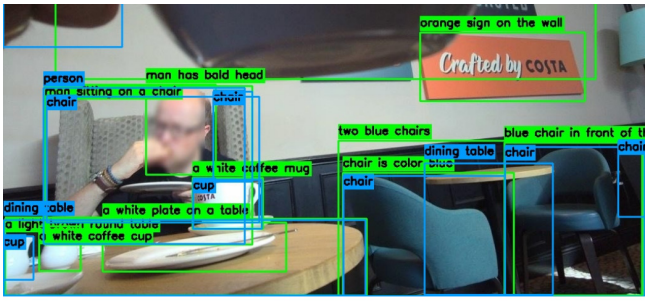


Types of annotations	MEMORIA
Object detection	[(tv, 0.471), (couch, 0.906), (cup, 0.432), (cup, 0.353), (bottle, 0.922)]
OCR	[('flora', 0.836),('10o', 0.387), ('gooune', 0.063)]
Caption generation	"view from the kitchen into the living room."
Deep annotation	[("a bottle of water", 0.607), ("a white bucket on the floor", 0.73 ("a calendar page with scenery on it", 0.6803), ("a skylight in th ceiling", 0.699), ("a white and blue label on a refrigerator", 0.61 ("an electrical outlet on the wall", 0.595), ("a newspaper on th wall", 0.620), ("green cap on a water bottle", 0.5839)]

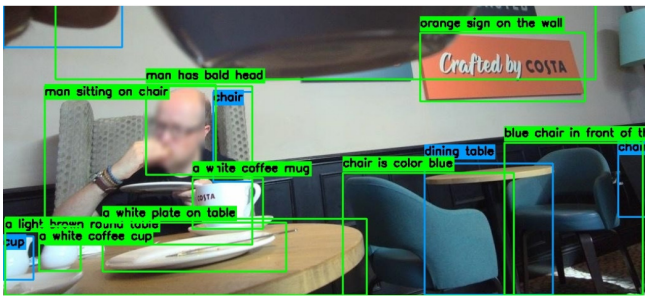
Types of annotations	MEMORIA
Object detection	[(person , 0.43), (car , 0.46)]
OCR	[("centra", 0.166)]
Caption generation	"driving a car in the rain."
Deep annotation	[("a man driving a car", 0.640), ("person in the car", 0.564), ("the hand of a person", 0.626)]

Types of annotations	MEMORIA
Object detection	[(car, 0.963), (car, 0.916), (car, 0.298)]
OCR	[('o6dl', 0.258), ('58377', 0.433)]
Caption generation	"property image # detached holiday home in quiet locatio with garden"
Deep annotation	[("a blue car parked on the street", 0.702), ("black car parked i street", 0.670), ("license plate on the car",0.707), ("red tail ligh on a car", 0.582), ("back tire of the car", 0.614), ("the back window of the car", 0.589), ("window on a brick building", 0.606 ("window on a building", 0.626), ("window on a building", 0.587 ("cars parked in front of a building", 0.609), ("a red brick building", 0.551), ("window of a blue van", 0.585), ("window in brick building", 0.645), ("window on a building", 0.622)]

Figure 3: Some examples of image annotations in MEMORIA.



(a) Object detections before reducing redundancy



(b) Object detections after reducing redundancy

Figure 4: In a), the original annotations obtained by the algorithms integrated into MEMORIA. In b), the impact of reducing redundancy between object detections generated by YOLOv7 (blue) and GRiT (green).

an event segmentation method based on hierarchical events to segment lifelog images into distinct events. The hierarchical method has 5 main layers:

- **Days:** This layer segments events based on daily life; a day is considered to begin when a person wakes up and ends when the person goes to sleep.
- **Parts of the Day:** Events are then segmented into multiple parts of the day which include morning, afternoon, and evening and their sub-categories (early and late).
- **Locations:** In this layer, events are segmented into locations and trajectories, with a trajectory being the data in the gap between one location and the next one. Locations are provided via the clustering method previously described in subsection 3.1
- **Environment:** In this layer events are segmented based on scene recognition described in 3.2.3. While in a location, a person may be indoors but then step outdoors; this layer is used to separate those 2 events.
- **Images Similarity:** This is the last main layer and segments events by sequential image similarity

In addition to the five main layers, it has two sub-layers, one between the Days and Parts of the Day layers, and the other is incorporated into every other layer. The first one applies Blind Image Quality Assessment (BIQA) to filter out low-quality images as they are not useful; the second sub-layer is used to fine-tune the boundaries of events in each layer. Figure 5 presents a diagram showcasing all layers of the event segmentation method.

3.4 Search Engine

MEMORIA can receive queries in the form of free text or with categorization filters. Free text search allows for a more flexible way of querying and exploring lifelog data. When querying using free text, the query is broken down into tokens and temporal keywords. Tokens can be objects, events, activities, locations, etc, and temporal keywords are words such as “after” and “before”. Tokens are then expanded using word expansion with the word2vec [10] technique, which is followed by the formation of the database query. MEMORIA uses a PostgreSQL database with multiple relationships between tables to store user data and processed data.

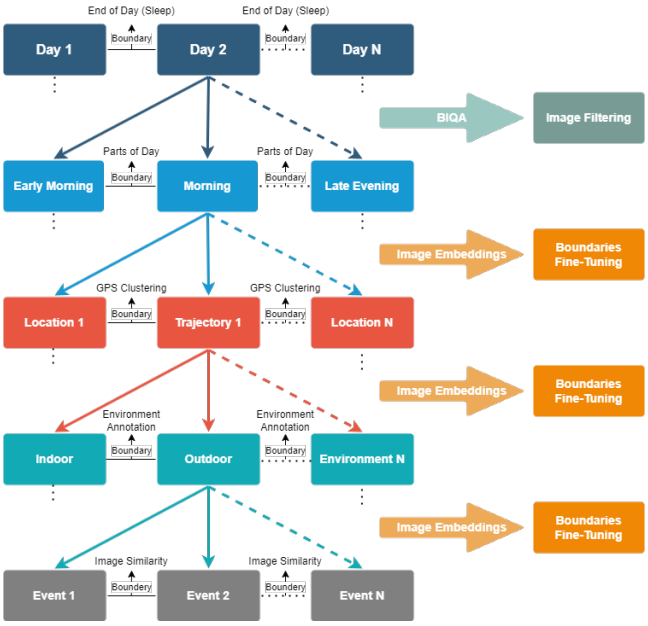


Figure 5: Representation of the hierarchical event segmentation algorithm.

4 RESULTS

By combining all the annotations generated by all the different methods, namely object detection, object understanding, OCR, scene understanding, and caption generation models, MEMORIA can, for each image, generate vast and rich amounts of annotations. Then, using the search engine, the user can search for specific moments using either free-text search or by using the filters.

This section will showcase results on the image annotation system of MEMORIA and provide an example of user interaction.

4.1 Image Annotation Results

This subsection will present results on the image annotation system of MEMORIA, with some insights into the lifelogger who captured the LSC dataset[6].

MEMORIA has processed the entire 18-month LSC dataset and annotated all images. Table 2 shows the annotations of each part of MEMORIA for the image in Figure 4. These annotations are then divided into categories so the user can use the filters when performing a query. The categories are:

- **Places**: Populated with the results from the *Categories* annotations of the Scene Understanding Model;
- **Objects**: This category is populated with the annotations from YOLOv7 and GRiT;
- **Attributes**: The results from the *Attributes* annotations of the Scene Understanding Model;
- **OCR**: This category contains all the text detected by CRAFT.

The LSC dataset has more than seven hundred thousand images, and MEMORIA extracted more than 15 million annotations. These are enough to get some insights into the lifelogger's life, an associate professor at Dublin City University's School of Computing

Table 2: Annotations extracted from a lifelog

Model	Annotations
YOLOv7	"cup": 0.78, "chair": 0.62, "chair": 0.61, "dining table": 0.52, "person": 0.32
GRiT	"black pot on hook": 0.72, "man sitting on chair": 0.67, "orange sign on wall": 0.73, "white coffee mug": 0.77, "chair is color blue": 0.61, "blue chair in front of desk": 0.62, "light brown round table", 0.64, "man has bald head": 0.59, "white plate on table": 0.59, "white coffee cup": 0.60
CRAFT	"costa": 0.68, "crafted": 0.58, "asted": 0.83, "by": 0.96, "costa": 0.91
ResNet 18	Attributes: {"cloth", "congregate", "enclosed area", "indoor lighting ", "man-made", "no horizon", "socialize", "work"} Categories: "categories": {"classroom": 0.043, "coffee shop": 0.33, "food court": 0.05, "restaurant": 0.08, "sushi bar": 0.05}
ClipCap	"A man sitting at a table with a coffee cup in front of him.", "This is a picture of a man sitting at a table in a coffee shop and drinking a cup of coffee."

in Ireland, or "DCU". Figures 6 and 7 are word clouds made of the annotations of YOLOv7 and GRiT, respectively.

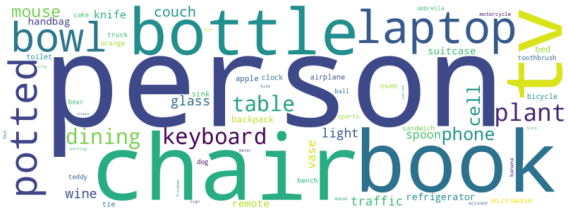


Figure 6: Word cloud of YOLOv7 annotations extracted from the LSC dataset.

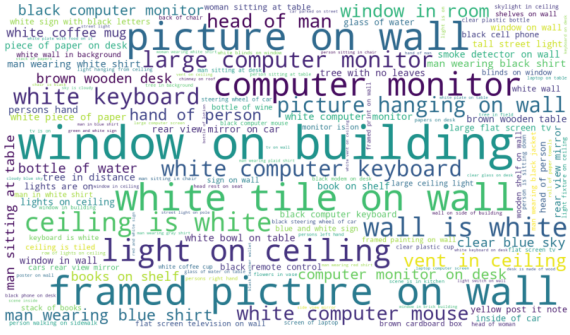


Figure 7: Word cloud of GRiT annotations extracted from the LSC dataset.

We can see from the word cloud of YOLOv7 that the most common annotations are objects that are very common in our everyday

lives, such as “tv” or “bottle”, however, we can also find annotations that can represent someone that works in an office like “laptop” and “keyboard”.

From the GriT annotations, we can also get annotations representing an office like “white computer keyboard” and “computer monitor”. However, the GriT annotations are especially useful for describing spatial relationships, like “framed picture on wall”, which add an extra layer of detail to the overall annotations.

Figure 8 shows the most common annotations from CRAFT, which is the OCR model used in MEMORIA.

Figure 8: Word cloud of OCR annotations extracted from the LSC dataset.

The annotations extracted from CRAFT allow us to infer that one common word is “costa”, which actually refers to the sign on Figure 4 which appears in a coffee shop that the lifelogger frequents. The word “a320” is also very common and refers to the airplane model Airbus A320, which if queried, we can verify that the lifelogger traveled multiple times on that aircraft model.

The final word cloud, shown in Figure 9 shows the annotations from the Scene Understanding model, trained on the Places365 dataset.

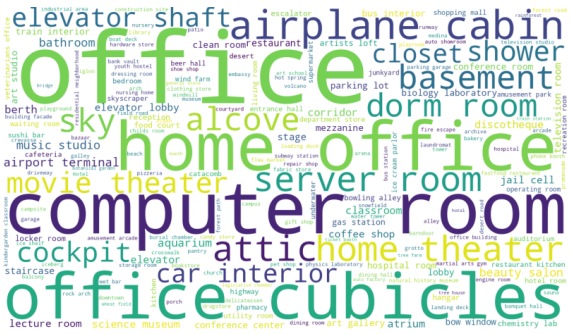


Figure 9: Word cloud of Places365 annotations extracted from the LSC dataset.

Based on these annotations, we can have an insight into the activities of the lifelogger. Annotations like “computer room”, “office” and “server room” are closely related to the lifelogger’s profession as an Associate Professor at DCU. These annotations, generated by the scene understanding model, are important because when users query MEMORIA, they usually search for images captured in specific environments, such as their office or home. Therefore, these annotations play a crucial role in enhancing the overall performance of MEMORIA.

4.2 Interaction Example

This subsection will present an interaction example using the following LSAT topic of the NTCIR17-Lifelog:

id: 17005
type: adhoc
uid: u1
title: Car stopping
description: Find examples of when I’m in the car stopping at a red traffic light.
narrative: To be relevant, these images should accurately depict the familiar scene of a stationary car awaiting the green light, capturing the essence of these routine pauses during my travels.

The keywords were extracted from the topic and introduced into MEMORIA. Specifically, the keywords used for searching were “red traffic light” and “car”. Once the search button is activated, the resulting images are shown to the user. Figure 10 shows this part of the process.

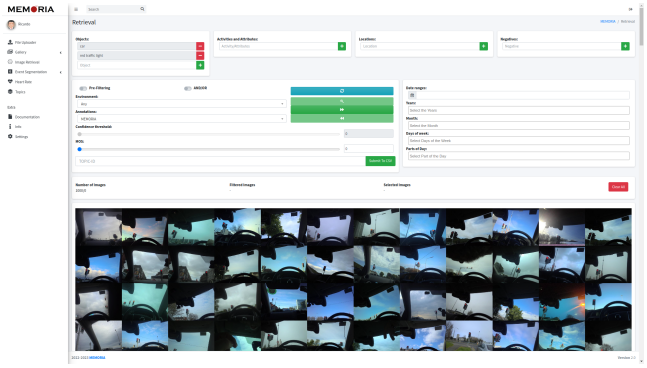


Figure 10: The retrieval user interface with the resulting images of a query.

In the search results, the user has the option to select either a single image or multiple images selecting the image that can be relevant to this topic. Figure 11 presents an example of the process of selecting multiple images of interest. In addition, images can be individually zoomed in by the user, as shown in figure 12.

The user can also retrieve images from 1 hour before or after a specific image. Figure 13 shows the images that happened 1 hour after the selected image. In this instance, after stopping at the red traffic light, the lifelogger proceeded to go shopping.

5 CONCLUSION

In this paper, we presented a lifelog retrieval system, denoted MEMORIA (Memory Enhancement and MOment RetrIeval Application), with the intent to participate in the NTCIR lifelog-5 event. The main goal of MEMORIA’s participation in this event is to evaluate the system’s performance and accuracy, with the

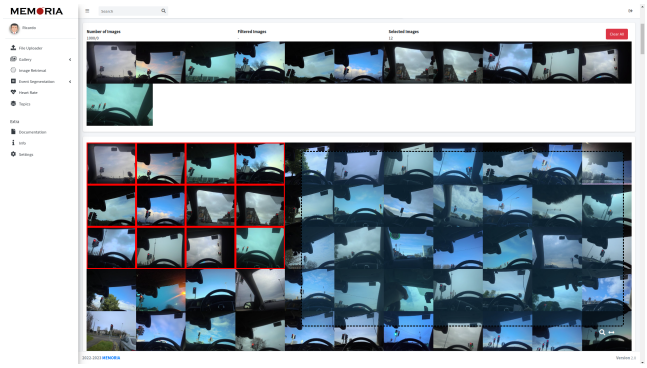


Figure 11: Example of selection of multiple images from the results.

Figure 12: Zoom in on a selected image

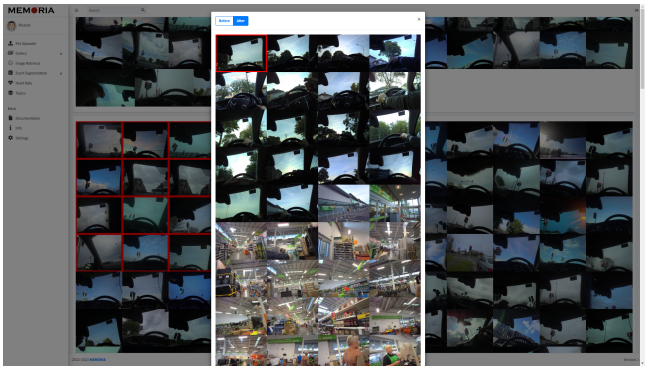


Figure 13: Images temporally acquired by the lifeloger after the selected image.

aim of gaining some insights into aspects on which the system can improve. This, in turn, will serve to advance the research on lifelog retrieval systems.

As future work, it would be interesting to investigate whether computer vision models trained on lifelog image datasets could yield more robust annotations. Furthermore, bridging the gap between queries submitted to a lifelogging system and the corresponding annotations could be achieved through research focused on understanding how humans search for their memories and how annotations can follow that structure.

The layers of the hierarchical segmentation approach can be extended into a different structure based on new and more fine-grained annotations. Additionally, the exploration of natural language processing techniques to generate descriptive textual representations of events within a sequence of images can be performed. This enhancement will enable lifelogging systems to query events based on textual descriptions, further improving retrieval capabilities. Moreover, using keywords like ”after” or ”before” will be processed autonomously by the system since the events are now tied to a chronological order.

Furthermore, the system is not taking full advantage of the event segmentation in the process of selecting the images to submit to the competition's server. Work is in progress to devise a more efficient way to select the events to be submitted.

ACKNOWLEDGMENTS

This work was supported in part by National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020, co-financed by COMPETE 2020 and Regional Operational Program Lisboa 2020, through Portugal 2020 and FEDER.

REFERENCES

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2023. Memento 3.0: An Enhanced Lifelog Search Engine for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 41–46. https://doi.org/10.1145/3592573.3593103

[2] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2023. Voxento 4.0: A More Flexible Visualisation and Control for Lifelogs *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 7–12. https://doi.org/10.1145/3592573.3593097

[3] et al. Baek, Jeonghun. 2019. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. https://doi.org/10.1109/ICCV.2019.00481

[4] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection. https://doi.org/10.48550/ARXIV.1904.01941

[5] Marc Bolanos, Mariella Dimiccoli, and Petia Radeva. 2016. Toward storytelling from visual lifelogging: An overview. *IEEE Transactions on Human-Machine Systems* 47, 1 (2016), 77–90.

[6] Cathal Gurrin, Björn Þór Jónsson, Duc Tien Dang Nguyen, Graham Healy, Jakub Lokoc, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, Werner Bailer, and Klaus Schoeffmann. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23 *(ICMR '23)*. Association for Computing Machinery, 678–679. https://doi.org/10.1145/3591106.3592304

[7] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:246411402

[8] et al. Lin, Tsung-Yi. 2014. https://doi.org/10.48550/ARXIV.1405.0312

[9] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2, 11 (2017), 205.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/ARXIV.1301.3781

[11] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. https://doi.org/10.48550/ARXIV.2111.09734

[12] 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. https://doi.org/10.48550/ARXIV.2103.00020

[13] Ricardo Ribeiro, Luísa Amaral, Wei Ye, Alina Trifan, António JR Neves, and Pedro Iglésias. 2023. MEMORIA: A Memory Enhancement and MOment RetrIeval Application for LSC 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*. 18–23.

[14] Ricardo Ribeiro, Alina Trifan, António JR Neves, et al. 2022. Lifelog Retrieval From Daily Digital Data: Narrative Review. *JMIR mHealth and uHealth* 10, 5 (2022), e30517.

[15] Klaus Schoeffmann. 2023. LifeXplore at the Lifelog Search Challenge 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 53–58. https://doi.org/10.1145/3592573.3593105

[16] Ly Duyen Tran, Binh Nguyen, Liting Zhou, and Cathal Gurrin. 2023. MyEachtra: Event-Based Interactive Lifelog Retrieval System for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 24–29. https://doi.org/10.1145/3592573.3593100

[17] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2020. Myscéal: An Experimental Interactive Lifelog Retrieval System for LSC'20. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge* (Dublin, Ireland) *(LSC '20)*. Association for Computing Machinery, New York, NY, USA, 23–28. https://doi.org/10.1145/3379172.3391719

[18] Ly-Duyen Tran, Dongyun Nie, Liting Zhou, Binh Nguyen, and Cathal Gurrin. 2023. VAISL: Visual-Aware Identification Of Semantic Locations In Lifelog. Springer-Verlag, Berlin, Heidelberg, 659–670. https://doi.org/10.1007/978-3-031-27818-1_54

[19] Quang-Linh Tran, Ly-Duyen Tran, Binh Nguyen, and Cathal Gurrin. 2023. MemoriEase: An Interactive Lifelog Retrieval System for LSC'23. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge* (Thessaloniki, Greece) *(LSC '23)*. Association for Computing Machinery, New York, NY, USA, 30–35. https://doi.org/10.1145/3592573.3593101

[20] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2022. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. https://doi.org/10.48550/ARXIV.2207.02696

[21] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. 2022. GRiT: A Generative Region-to-text Transformer for Object Understanding. *arXiv preprint arXiv:2212.00280* (2022).

[22] Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, and Quannan Li. 2011. Geolife GPS trajectory dataset - User Guide. *Geolife GPS trajectories* (July 2011).

[23] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

[24] Liting Zhou, Cathal Gurrin, Duc-Tien Dang-Nguyen, Graham Healy, ChenYang Lyu, Tianbo Ji, Longyue Wang, Hideo Joho, Ly-Duyen Tran, and Naushad Alam. 2023. Overview of the NTCIR-17 Lifelog-5 Task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-17)*. Tokyo, Japan.