National College of Ireland

A Deep Learning Framework for Memory Retrieval from Lifelogging Data

MSc Research Project
Data Analytics

Mohammad Aman
Student ID: X21109079

School of Computing
National College of Ireland

Supervisor: Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Mohammad Aman
Student ID:	X21109079
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Paul Stynes
Submission Due Date:	15/12/2022
Project Title:	A Deep Learning Framework for Memory Retrieval from Lifelogging Data
Word Count:	2830
Page Count:	11

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deep Learning Framework for Memory Retrieval from Lifelogging Data

Mohammad Aman
X21109079

Abstract

An emerging trend known as lifelogging is a process of digitally documenting and processing the data of an individual's daily experiences. Lifelogging creates data which can be noisy and with continuity; therefore, it is challenging to give a comprehensive means of retrieving events or moments of interest to the public. This research proposes a deep learning framework to improve memory retrieval from lifelogging data. The proposed framework combines text-image embeddings and ensembles of a zero-shot deep learning model. The framework is implemented using three versions of the Contrastive Language-Image Pre-training (CLIP) model based on the combination of 12 datasets created by seven users containing more than 100000 images. The results are evaluated based on the average precision@k metric for different values of k. This framework improves retrieval performance and shows the possibility of helping people who have Alzheimer's and other forms of dementia to recall useful information using the retrieval framework.

1 Introduction

Lifelogging allows people to digitally document their lives using a wearable camera, which can be used to characterize their everyday routines, emotions, and interactions. Lifelogging captures egocentric data, which can be leveraged to understand contextual information about an individual's lifestyle by retrieving specific moments (Gurrin et al.; 2014). One of the critical challenges of lifelog retrieval is bringing back recent memories, especially in people with Alzheimer's or other types of dementia (Carós et al.; 2020). However, analyzing egocentric data is challenging due to its size and thousands of images daily. For instance, developing a lifelog retrieval system that people with memory impairment can use personally and reduce the need to take memory improvement sessions.

The aim of the research is to improve the performance of a memory retrieval system based on lifelogging images by making use of deep learning techniques. The research investigates to what extent a deep learning framework can improve the accuracy of memory retrieval from lifelogging data to help people with Alzheimer's and other forms of dementia. The major contribution of this research is a deep learning framework that uses pre-trained deep learning models and ensemble methods that improve the accuracy of an image retrieval system from lifelogging data. Another contribution is a notebook interface to input text queries and retrieve images. To address the research question, the following specific sets of research objectives were derived:

- Investigate state of the art broadly around current deep learning techniques and models for information retrieval from egocentric data.
- Design a deep learning framework that combines enhanced image embeddings using deep learning techniques to improve search quality and retrieval of images from lifelogging data.
- Implement a deep learning framework using ensembles of three (CLIP) models. The three models, RN50x64 (He et al.; 2016), ViT-L/14 (Dosovitskiy et al.; 2020) and ViT-L/14@336px (Radford et al.; 2021), are used to encode the semantic representation of images and generate rankings for high–dimensional image embeddings. These rankings are tested to reduce the semantic gap between images and queries to outperform other models.
- Evaluate the framework's performance based on the accuracy metric, such as average precision@k for evaluation queries.

Section 2 related work discusses the progress of deep learning models used in lifelogging data. Section 3 describes the suggested methodology, briefly explains the dataset, and discusses the proposed techniques to carry out the project. In the next section, the implementation of this research is discussed, followed by the results.

2 Related Work

Lifelogging is a process of digitally documenting and processing the data of an individual's daily experiences. Lifelogging allows capturing egocentric images that show the activities of the lifelogger from a first-person perspective. Lifelogs provide useful source of information on human lives and memory. However, the analysis of lifelog collections is challenging due to their size (thousands of images every day) (Talavera et al.; 2020).

It has recently gained popularity due to the low cost of computer storage, willingness to store and share our data, and advancement in sensory devices and technology like Google Glass has made lifelogging a matter of public interest (Gurrin et al.; 2014). However, several challenges must be addressed, from enhanced data capturing to accurate information retrieval (Gurrin et al.; 2008; Allan et al.; 2012).

Several applications emerged from lifelogging, from quantified self-analytics to applications in healthcare and sports (Amin et al.; 2016)(Barua et al.; 2013). One such application of lifelog retrieval is developing a mechanism to bring back recent memories by acting as a memory prosthetic, for example, to help people with Alzheimer's' and other types of dementia (Bahrainian and Crestani; 2016). The authors proposed a system for compiling social interactions of a user by associating subjects with the relevant image, place, and time. (Tancharoen and Aizawa; 2004) developed one of the earliest context-based lifelog retrieval systems using the wearable camera. Since then, several studies have analysed lifelogging data from food-images categorization (Martinez et al.; 2019), identifying activities (Furnari et al.; 2016a, 2015, 2016b), and sentiment analysis (Talavera et al.; 2017) to human interactions analysis (Aghaei et al.; 2018; Alletto et al.; 2015). LEMoRe (deOliveira Barra et al.; 2015) was one of the early moment retrieval systems from a lifelog with a user-friendly query interface facilitating improved semantic access. The semantic information using objects and tags was considered using a heatmap, and a tag suggestion strategy was used for novice users.

Mysceal (Tran et al.; 2020) utilized a new algorithm "aTFIDF" for matching a query with image annotations to rank document retrieval. The system (Tran et al.; 2020) leveraged improved image retrieval and an interactive search engine with query suggestion. Commonly, embedding approaches are based on encoding concepts from image labels and queries into the same vector space to calculate their similarity (Gong et al.; 2014; Karpathy and Fei-Fei; 2015; Hodosh et al.; 2013). LifeGraph (Rossetto et al.; 2020) used a knowledge graph structure with nodes indicating things or scenes detected in photos. These entities can be paired with matching images and objects to enhance the knowledge base with related objects and activities. Some systems improved by encoding the whole textual input or constructing captions for lifelog that characterize the events represented in them (Chu et al.; 2020).

Liting et al. (Zhou and Gurrin; 2022) recently proposed effective image annotation using a pre-trained image captioning model. The first approach captured the user's intentions using semantic correlation, while the following approach represented visual and text semantics into a common latent space using embeddings which proved to be a powerful technique in image annotation. The findings show new possibilities for research into text and multimodal-embedding models.

A recently introduced lifelog retrieval system (Alam et al.; 2022) also incorporated high dimensional image emeddings for image retrieval. The system showed promising results using the pre-trained zero-shot models. However, the system was implemented on the specific dataset and used same evaluation queries explicitly designed for the considered dataset.

In conclusion, the state-of-the-art reveals that using pre-trained models such as CLIP to derive embeddings with ensemble techniques has improved the lifelog retrieval systems. There is an experimental need to determine an optimal model for lifelog retrieval. State-of-the-art indicates that the research on lifelog data is limited to specific datasets, for example, Lifelog Search Challenge (LSC) dataset (Gurrin et al.; 2022), which is not publicly available. However, there is a need to expand the research on different lifelogging datasets to examine the effectiveness of the models such as CLIP on retrieval systems. This study presents a deep learning framework that combines multiple versions of the deep learning CLIP model with ensemble techniques to improve memory retrieval from lifelogging data. An optimal model will be identified by generating rankings with high-dimensional image embeddings and testing retrieval results using natural language queries.

3 Methodology

The research methodology consists of the following steps as shown in Fig. 1. The first step, Data Collection involves gathering and combining EgoRoutine datasets (Talavera et al.; 2020). The data is created using a wearable camera by seven users over 104 days. The dataset size is 57.3 gigabytes that contain more than 100000 images.

The second step, Data Pre-processing, involves organizing and cleaning the dataset. Lifelogging images are captured at short intervals (2fpm in considered dataset), which stores redundant and blurred images. All the images were individually examined and blank images were removed. After pre-processing, the dataset contained 98000 images and was merged into a single folder and uploaded to a cloud server.

The third step, Data Modelling, involves model deployment, image and text data

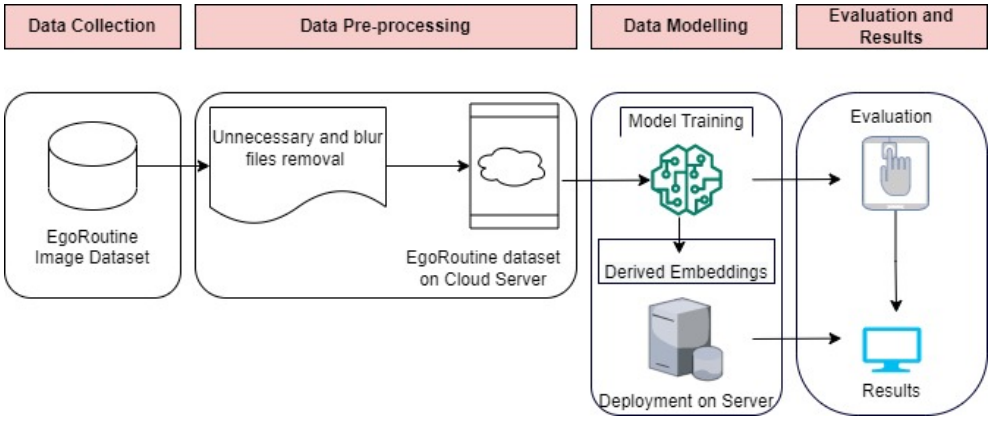

Figure 1: Research Methodology

projection into a common latent space to derive embeddings from the EgoRoutine dataset. Three versions of the state-of-the-art CLIP model with zero-shot capabilities namely, RN50x64, ViT-L/14 and ViT-L/14@336px were used. These models were trained on over four hundred million images to discover visual concepts that can be applied to activities, such as optical character recognition, object detection, image retrieval, etc. Ensemble techniques were employed to capture improved embeddings using multiple models. A weighted sum of scores from the three models RN50x64, ViT-L/14 and ViT-L/14@336px were chosen in different ratios to optimise the output. The models, derived embeddings and other elements were migrated on a local device to query and retrieve images from lifelogging data.

The fourth step, Evaluation and Results, involves evaluating the performance of each of the deep learning models and their ensembles using the average precision@k metric. The accuracy of all the models was compared, and an optimal model was selected to perform memory retrieval.

4 Design Specification

The architecture of the deep learning memory retrieval framework combines the EgoRoutine dataset and deep learning image classification models as shown in Fig. 2. The first component of the framework includes image data and high-dimensional image embeddings as discussed in section 4.1. The components of the retrieval framework are discussed in section 4.2.

4.1 Deep Learning Lifelog Classification Model

The memory retrieval framework starts by loading the egocentric data into the model and encoding image batches to compute feature vectors. The derived embeddings are stored on the server in a NumPy file. The indices are stored in a separate photo_indices file for the model to match and retrieve images. The framework employs three CLIP models namely, RN50x64, ViT-L/14 and ViT-L/14@336px, to classify and rank all images based on the cosine similarity.

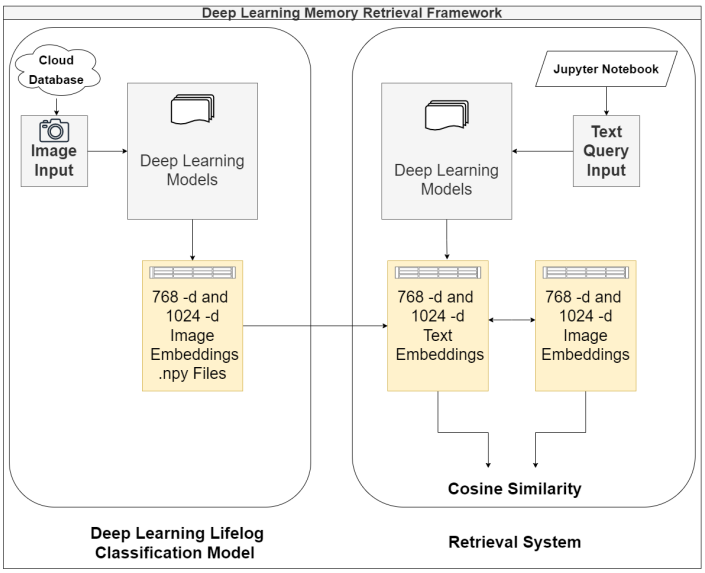

Figure 2: Deep Learning Framework Architecture

4.2 Retrieval System

The retrieval part of the framework includes derived image and text embeddings, ensembles of three deep learning models and a query interface. The textual queries are passed through text encoders, and image-text embeddings are compared using cosine similarity to obtain a list of ranked image indices. In the retrieval phase, the user enters a search query to retrieve relevant images based on the final ranking. This user interaction is facilitated by migrating the server to the local host in jupyter notebook.

5 Implementation

The deep learning memory retrieval framework was implemented by connecting the code implementation on the server to the Jupyter notebook on the local machine. The AWS server (EC2-g4ad.4xlarge) with the Ubuntu platform was imported in Visual Studio Code and python language was used to implement the framework. FileZilla was used to connect the server and exchange data with the local machine. Trained image embeddings were stored as NumPy files on the server. An additional file with image indices was created to facilitate matching and quick retrieval of images. The text encoder of the model converts the textual query in its vector form to match image semantics. The model then compares the lifelogging dataset to compute cosine similarity and rank images. The ranked images are then retrieved and displayed on the screen, as shown in Fig. 3. The EgoRoutine

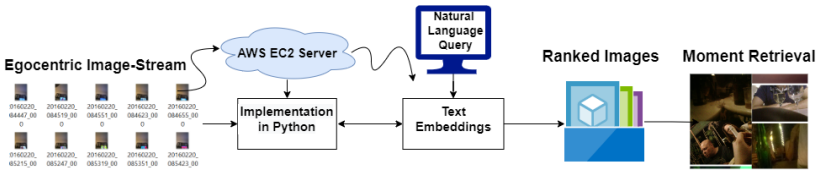

Figure 3: Implementation Flow

dataset was uploaded to the AWS EC2 server from the local disk using Filezilla. The required libraries, for example, PyTorch and clip, were installed through the terminal using the Putty tool. The CLIP models were trained with the data, and the embeddings were imported on the local machine. The .ipynb file was migrated from the server to the Jupyter notebook on the localhost. The dataset implementation is shown in Fig. 4.

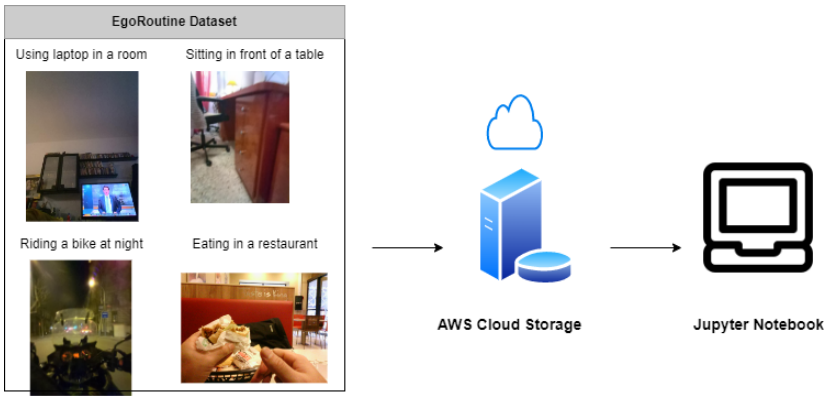

Figure 4: EgoRoutine Implementation

6 Results and Evaluation

The aim of this experiment is to compare and evaluate the accuracy of three deep learning models on seven EgoRoutine datasets. Three deep learning CLIP models, namely, RN50x64, ViT-L/14 and ViT-L/14@336px and their ensemble models are used to derive embeddings and retrieve lifelogging images based on ten natural language queries. The three models and their ensembles are compared based on precision, as shown in Table 7. We tested our framework based on the visual data available to us using precision@k. We evaluated the performance based on 10 natural language queries, for example, "I was attending a lecture and there was a presentation on projector". The backend model lacks a sense of explicit information like time, date, etc. The queries are evaluated based on accurate retrieval using multiple k values. Based on the dataset, the memory retrieval framework can handle complicated queries efficiently, e.g. a query looking for cliffs and sea with a beautiful blue sky and good weather.

We evaluate the models on Precision@k metric which is used to find only target images among the k images in the output. Table 7 shows average precision from all the models at different values of k.

Table 1: Precision@k calculated at multiple k values for all the models along 10 queries.

Model	**k=5**	**k=10**	**k=20**	**k=30**	**k=50**
ResNet50x64	**0.74**	0.71	0.705	0.703	0.672
ViT-L/14	0.82	**0.84**	0.79	0.76	0.722
ViT-L/14@336px	0.84	**0.86**	0.81	0.77	0.774
Ensemble 1	**0.88**	0.85	0.825	0.803	0.756
Ensemble 2	**0.90**	0.87	0.84	0.833	0.81
Ensemble 3	0.84	**0.86**	0.81	0.78	0.756

6.1 Experiment 1

ResNet50x64: A model using 64x the compute of previous model ResNet-50 and generates 1024 dimensional image-text embeddings.

Table 2: Precision@k calculated for ResNet50x64.

Model	**k=5**	**k=10**	**k=20**	**k=30**	**k=50**
ResNet50x64	0.74	0.71	0.705	0.703	0.672

6.2 Experiment 2

ViT-L/14: A Vision Transformer model that generates 768 dimensional image-text embeddings.

Table 3: Precision@k calculated for ViT-L/14.

Model	**k=5**	**k=10**	**k=20**	**k=30**	**k=50**
ViT-L/14	0.82	0.84	0.79	0.76	0.722

6.3 Experiment 3

ViT-L/14@336px: Most recent version of the CLIP model which outperformed other models in zero-shot prediction.

Table 4: Precision@k calculated for ViT-L/14@336px.

Model	**k=5**	**k=10**	**k=20**	**k=30**	**k=50**
ViT-L/14@336px	0.84	0.86	0.81	0.77	0.774

6.4 Experiment 4

Ensemble of ResNet50x64 and ViT-L/14: Cosine scores are taken as weighted sum in 1:1 ratio.

Table 5: Precision@k calculated for ensemble of RN50x64 & ViT-L/14@336px (1:1).

Model	**k=5**	**k=10**	**k=20**	**k=30**	**k=50**
Ensemble 1	0.88	0.85	0.825	0.803	0.756

6.5 Experiment 5

Ensemble of ResNet50x64 & ViT-L/14: Cosine scores are taken as weighted sum in 3:1 ratio.

Table 6: Precision@k calculated for ensemble of RN50x64 & ViT-L/14@336px (3:1).

Model	**k=5**	**k=10**	**k=20**	**k=30**	**k=50**
Ensemble 2	0.90	0.87	0.84	0.833	0.81

6.6 Experiment 6

Ensemble of ViT-L/14 & ViT-L/14@336px: Cosine scores are taken as weighted sum in 3:1 ratio.

Table 7: Precision@k calculated for ensemble of ViT-L/14 & ViT-L/14@336px (3:1).

Model	k=5	k=10	k=20	k=30	k=50
Ensemble 3	0.84	0.86	0.81	0.78	0.756

6.7 Discussion

The system intends to include better image embeddings to optimize and improve the retrieval results from the lifelogging data. Three CLIP models and their ensembles are incorporated into the proposed system. The results indicate that the most recent CLIP version ViT-L/14@336px (Radford et al.; 2021) performs better on the data than the previously released versions, namely, ResNet50x64 (He et al.; 2016) and ViT-L/14 (Dosovitskiy et al.; 2020), when compared individually. Further, we implement the framework using ensembles of the three models discussed above. We observe that the ensemble of ResNet50x64 and ViT-L/14 in the ratio 3:1 outperforms all other models with considerable high precision. The highest precision is found to be 0.90 at $k = 5$.

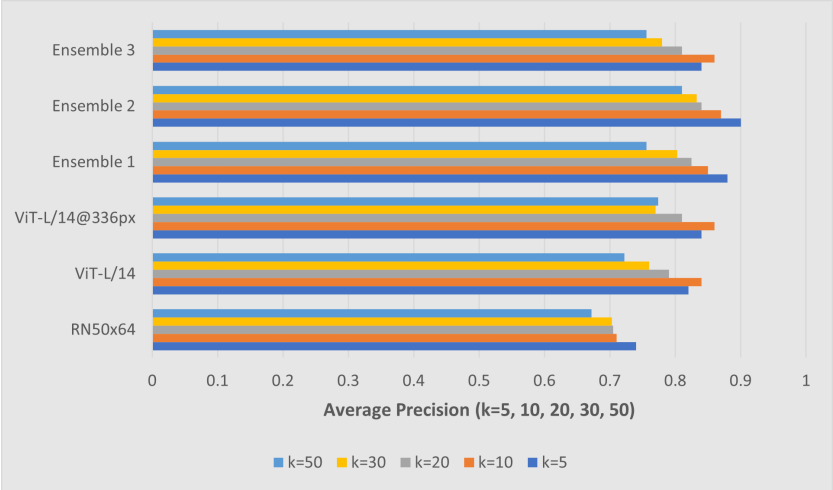

Figure 5: Comparison of average precision@k calculated for all six models.

Figure 5 compares the accuracy of each of the six models. The characteristics of the memory retrieval framework indicate that the value of k is important to consider. The average precision at lower k values e.g. 5 and 10 is higher which decreases as k increases. The models work efficiently fast and retrieve results within 3 seconds after running the query.

The current lifelog retrieval systems are majorly evaluated on intentionally tailored textual queries based on the specific datasets. The proposed framework works well on the random natural language queries to recall past memory and identify specific moments.

7 Conclusion and Future Work

The aim of this research was to present an improved memory retrieval system based on lifelogging data using deep learning techniques. This research proposes a deep learning framework that combines text-image embeddings and ensembles of a zero-shot deep learning CLIP models. Results demonstrate that the most recent version of the model adapts well on the dataset and give more accurate results. Through the experiments, it is shown that the ensemble methods can leverage enhanced image embeddings which further improves the retrieval accuracy. The ensemble of CLIP models ResNet50x64 and ViT-L/14 in the ratio 3 : 1 outperforms all the other models and give highest precision in retrieval accuracy. A limitation of this study is the insufficient details in the dataset, for example, temporal data. With more information on date, location and time, the framework can be used to retrieve more precise results.

This research encourages researchers which are confined to improve the lifelog retrieval systems using the same dataset, to develop a framework performing well on different datasets. The research can potentially enhance the memory retrieval systems with interactive user interface. This work can be improved by optimizing the models with more experiments on ensembles and selecting a model with high accuracy on large number of image retrievals. Based on this framework, an extensive research can be done in the field of healthcare for the patients with memory impairment with the availability of personal lifelogs. An easy-to-use user-interface and improved datasets will enable people to use the system and recall specific moments from past, which can reduce the constant need to take memory improvement sessions. For example, a person recalling a particular moment just after he purchased a coffee on a specific date or developing a real-time system that can learn and recall simultaneously. More study needs to be done for retrievals in such systems to identify an optimal memory retrieval model.

References

Aghaei, M., Dimiccoli, M., Ferrer, C. C. and Radeva, P. (2018). Towards social pattern characterization in egocentric photo-streams, *Computer Vision and Image Understanding* **171**: 104–117.

Alam, N., Graham, Y. and Gurrin, C. (2022). Memento 2.0: An improved lifelog search engine for lsc'22, *Proceedings of the 5th Annual on Lifelog Search Challenge*, pp. 2–7.

Allan, J., Croft, B., Moffat, A. and Sanderson, M. (2012). Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne, *Acm sigir forum*, Vol. 46, ACM New York, NY, USA, pp. 2–32.

Alletto, S., Serra, G., Calderara, S. and Cucchiara, R. (2015). Understanding social relationships in egocentric vision, *Pattern Recognition* **48**(12): 4082–4096.

Amin, M. B., Banos, O., Khan, W. A., Muhammad Bilal, H. S., Gong, J., Bui, D.-M., Cho, S. H., Hussain, S., Ali, T., Akhtar, U. et al. (2016). On curating multimodal sensory data for health and wellness platforms, *Sensors* **16**(7): 980.

Bahrainian, S. A. and Crestani, F. (2016). Cued retrieval of personal memories of social interactions.

Barua, D., Kay, J. and Paris, C. (2013). Viewing and controlling personal sensor data: what do users want?, *International Conference on Persuasive Technology*, Springer, pp. 15–26.

Carós, M., Garolera, M., Radeva, P. and Giro-i Nieto, X. (2020). Automatic reminiscence therapy for dementia, *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pp. 383–387.

Chu, T.-T., Chang, C.-C., Yen, A.-Z., Huang, H.-H. and Chen, H.-H. (2020). Multimodal retrieval through relations between subjects and objects in lifelog images, *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*, pp. 51–55.

deOliveira Barra, G., i Nieto, X. G., Cartas-Ayala, A. and Radeva, P. (2015). Lemore: A lifelog engine for moments retrieval at the ntcir-lifelog lsat task, *The 12th NTCIR Conference, Evaluation of Information Access Technologies* .

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* .

Furnari, A., Farinella, G. M. and Battiato, S. (2015). Recognizing personal contexts from egocentric images, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1–9.

Furnari, A., Farinella, G. M. and Battiato, S. (2016a). Recognizing personal locations from egocentric videos, *IEEE Transactions on Human-Machine Systems* **47**(1): 6–18.

Furnari, A., Farinella, G. M. and Battiato, S. (2016b). Temporal segmentation of egocentric videos to highlight personal locations of interest, *European Conference on Computer Vision*, Springer, pp. 474–489.

Gong, Y., Ke, Q., Isard, M. and Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics, *International journal of computer vision* **106**(2): 210–233.

Gurrin, C., Byrne, D., O'Connor, N., Jones, G. J. and Smeaton, A. F. (2008). Architecture and challenges of maintaining a large-scale, context-aware human digital memory, *2008 5th International Conference on Visual Information Engineering (VIE 2008)*, IET, pp. 158–163.

Gurrin, C., Smeaton, A. F. and Doherty, A. R. (2014). Lifelogging: Personal big data.

Gurrin, C., Zhou, L., Healy, G., Jónsson, B. ., Dang-Nguyen, D.-T., Lokoč, J., Tran, M.-T., Hürst, W., Rossetto, L. and Schöffmann, K. (2022). Introduction to the fifth annual lifelog search challenge, lsc'22, *Proc. International Conference on Multimedia Retrieval (ICMR'22). ACM, Newark, NJ.*

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hodosh, M., Young, P. and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics, *Journal of Artificial Intelligence Research* **47**: 853–899.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137.

Martinez, E. T., Leyva-Vallina, M., Sarker, M. M. K., Puig, D., Petkov, N. and Radeva, P. (2019). Hierarchical approach to classify food scenes in egocentric photo-streams, *IEEE journal of biomedical and health informatics* **24**(3): 866–877.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision, *International Conference on Machine Learning*, PMLR, pp. 8748–8763.

Rossetto, L., Baumgartner, M., Ashena, N., Ruosch, F., Pernischová, R. and Bernstein, A. (2020). Lifegraph: a knowledge graph for lifelogs, *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*, pp. 13–17.

Talavera, E., Strisciuglio, N., Petkov, N., Radeva, P., Alexandre, L. A., Sanchez, J. S. and Rodrigues, J. M. (2017). Sentiment recognition in egocentric photostreams: Iberian conference on pattern recognition and image analysis, *8th Iberian Conference on Pattern Recognition and Image Analysis, IbPRIA 2017*, Springer, pp. 471–479.

Talavera, E., Wuerich, C., Petkov, N. and Radeva, P. (2020). Topic modelling for routine discovery from egocentric photo-streams, *Pattern Recognition* **104**: 107330.

Tancharoen, D. and Aizawa, K. (2004). Novel concept for video retrieval in life log application, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **3332**.

Tran, L.-D., Nguyen, M.-D., Binh, N. T., Lee, H. and Gurrin, C. (2020). Myscéal: an experimental interactive lifelog retrieval system for lsc'20, *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*, pp. 23–28.

Zhou, L. and Gurrin, C. (2022). Multimodal embedding for lifelog retrieval, Vol. 13141 LNCS.