Enhancing Lifelog Retrieval through Automatic Image Annotation and Computer Vision Techniques

Luísa Amaral
luisaamaral9@ua.pt
Ricardo Ribeiro
rfribeiro@ua.pt
António J. R Neves
an@ua.pt

IEETA, DETI, LASI
University of Aveiro
Aveiro, Portugal

Abstract

Lifelogging has recently gained significant popularity as individuals increasingly capture and document their daily experiences through different devices. This generates a vast digital collection of lifelogs, that hold valuable insights into the lifelogger's behaviors and patterns. Visual data is one of the most valuable sources of information in lifelogs and automatic image annotation plays a vital role in extracting information from lifelog images, enabling a comprehensive understanding of lifelog data and facilitating effective retrieval processes. This paper presents an exploration of different computer vision techniques that can be employed to extract annotations from lifelog images. Additionally, a practical implementation of a lifelog annotation pipeline by incorporating state-of-the-art techniques from diverse computer vision tasks is also described. Furthermore, this work introduces research efforts focused on reducing redundancy in annotations and presents the classification of their importance. The use of computer vision techniques allows the extraction of rich insights from lifelog images, leading to enhanced efficiency and accuracy when retrieving them. By leveraging these techniques, lifeloggers can gain deeper insights into their experiences, enabling a better understanding of their captured memories.

1 Introduction

Lifelogging is a process that involves actively or passively recording various aspects of an individual's daily activities using a range of digital recording devices. It has recently gained popularity due to the advancement of modern technology, and it aims to generate a comprehensive digital collection of personal records, called lifelogs, that capture various details of an individual's daily experiences. The value that lifelogs hold is immense since various information relating to an individual's activities, experiences, and behaviors can be extracted from them [6].

The most common lifelog that individuals capture daily is visual data in the form of images. Analyzing these images can provide valuable insights into an individual's daily experiences and memories. However, manual analysis of such a large volume of data can be time-consuming and impractical. To address this challenge, automatic image annotation algorithms can be implemented to process visual data. Enriching the information that can be extracted from lifelogs enables the retrieval of specific images when an individual wants to recollect a particular moment from their life.

2 Image annotation

Automatic image annotation is the process of automatically assigning tags and other metadata to digital images, aiming to close the gap between textual queries and images.

The complexity of lifelog images can range widely, from simple images with minimal activity to highly complex scenes. Given this variability of content, it is important to consider a range of annotation techniques in order to effectively extract useful information from lifelog images in multiple levels of detail. On a lower level, object detection can be used to identify specific objects within the image. Object understanding adds to that information by generating a descriptive sentence for every identified object. With the aid of optical character recognition (OCR), text can be extracted from images, being particularly useful for capturing textual information that an individual encounters throughout their day. On a higher level, scene understanding can help identify the overall context of the image and the type of environment it depicts, providing contextual suggestions about the many locations that a person visited. Finally, automatic image captioning can be used to generate a textual description of the image, which can provide a high-level overview of its contents.

3 Proposal of an improved Annotation Pipeline for the MEMORIA lifelogging system

The aforementioned computer vision tasks were integrated within the lifelogging system MEMORIA [7], a system that stores lifelog images and allows their retrieval based on textual queries. The aim of this use case was to demonstrate the effectiveness and practicality of the proposed computer vision tasks in a lifelogging context using a real-world scenario.

The initial step consisted of updating the computer vision models that the system already included and adding newer models. An annotation processing stage was also added to minimize redundancy, evaluate the annotation's importance and detect object-OCR overlaps. This new annotation pipeline was employed to process the entire set of images from the LSC dataset [3], which consisted of an eighteen-month multimodal lifelog dataset captured between January 2019 and June 2020.

When combined with the annotation processing stage, the joint utilization of these computer vision models further amplifies the value of annotations. An example of an annotated lifelog is presented in Figure 1, along with the phases of the annotation pipeline.

3.1 Integration of computer vision models

The initial version of MEMORIA focused on extracting keywords and concepts from images using the object detection model YOLOv5 [1], along with a scene understanding model, ResNet18 architecture trained with the Places dataset [9]. To improve the system's annotation capabilities, additional models were introduced. YOLOv7 [8] replaced YOLOv5, improving detection accuracy. The GRiT model enhanced object understanding and provided detailed descriptions. CRAFT [2] was added to extract text from lifelogs. Higher-level annotations were improved with the addition of the ClipCap model [4], generating a caption for each image. The scene understanding model was retained, as it provided detailed information regarding the context of the scenes depicted in the lifelogs. These enhancements enriched the system's annotations and improved semantic matching with user queries.

3.2 Background and Foreground Distinction

Lifelog images tend to depict intricate scenes, with various objects and elements that are often organized in complex layouts, with challenging lighting conditions, blurriness, or clutter. To identify the most relevant objects in the context of the scene, depth maps were explored. These maps provide information about the distance of objects from the camera and hence enable the identification of objects that are closer to the lifelogger and consequently are more likely to be important for the understanding of the scene. A depth map was generated for each lifelog image using a machine learning model [5] that estimated a value of visual depth for each image pixel. The average of these values was then defined as a threshold. The identified objects that had a majority of pixels with a depth value higher than the threshold were considered to belong to the foreground plane, while the remaining to the background.

3.3 Annotation Redundancy Reduction

To identify and remove possible cases of repetition where the object detection and understanding models detect and classify the same object uti-

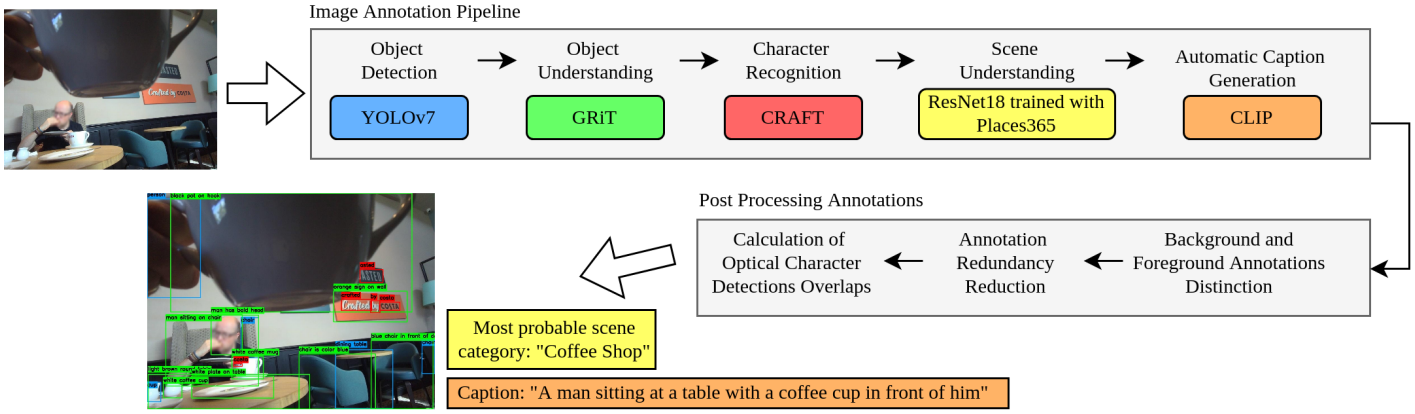

Figure 1: Lifelogging system MEMORIA's improved annotation pipeline.

lizing similar terms or concepts, the YOLO and GRiT detections are compared in pairs, and the ratio of overlap between the two bounding boxes of the pair is calculated. If the ratio exceeds 0.4, the assigned image planes for each object, as determined by the depth map, are compared. If the image planes match, it can be concluded that both annotations have most likely identified the same object. To further validate this conclusion, the similarity between the textual content of the detections is calculated.

Using the method of tokenization, the descriptive sentence generated by GRiT is split into words, and post-tagging is used to extract the nouns from the sentence. Each noun is then compared with the class that was assigned by YOLO, obtaining a similarity value for each comparison. This similarity can be computed taking into account the hierarchy that leads to the lowest common hypernym of the two words, which is the most general concept that both nouns share. Finally, a similarity score is computed on a scale of 0 to 1, where a higher score indicates a greater similarity in meaning between the two nouns. If a similarity score exceeds 0.5, it is inferred that both annotations redundantly describe or identify the same object. In such cases, only the most descriptive annotation, generated by GRiT, is retained as it typically provides more comprehensive details.

3.4 Optical Character Detection Processing

The outputs of the object detection models and of the OCR model can be combined in order to understand in what objects the detected text is written. If more than 50% of the pixels of an OCR detection overlap with an object detection, it can be said that the detected characters are written on that object.

4 Conclusion

This work explored the importance of automatic image annotation methods in the context of lifelogging and their potential to extract valuable information from lifelog images. Ultimately, one of the main challenges of a lifelogging system is to extract value from the collection of lifelog images to close the gap between queries and images. Using multiple annotation techniques can provide a more complete understanding of the lifelog data, enabling a more effective retrieval process and facilitating the identification of trends within the data. Furthermore, by extracting annotations on multiple levels, a higher level of flexibility is offered in the retrieval process.

The use case presented showcased the practical implementation of the selected models in a lifelog annotation pipeline. The annotation process was enhanced by these techniques and rich semantic information was extracted from lifelog images. An additional stage was also introduced in the annotation pipeline that effectively reduced redundancy between annotations, inferred their importance and determined the objects in which detected characters were written, improving the system's overall efficiency and accuracy.

Overall, integrating computer vision models into these systems holds immense potential for revolutionizing the way we capture, manage, and relive our personal experiences, taking a significant step towards a more intuitive and personalized lifelogging experience.

References

[1] ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, November 2022. URL https://zenodo.org/record/3908559.

[2] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character Region Awareness for Text Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9357–9366, Long Beach, CA, USA, June 2019. IEEE. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00959. URL https://ieeexplore.ieee.org/document/8953846/.

[3] Cathal Gurrin et al. Introduction to the sixth annual lifelog search challenge, lsc'23. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 678–679, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701788. doi: 10.1145/3591106.3592304. URL https://doi.org/10.1145/3591106.3592304.

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

[5] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168, Montreal, QC, Canada, October 2021. IEEE. ISBN 9781665428125. doi: 10.1109/ICCV48922.2021.01196. URL https://ieeexplore.ieee.org/document/9711226/.

[6] Ricardo Ribeiro, Alina Trifan, and António J R Neves. Lifelog Retrieval From Daily Digital Data: Narrative Review. *JMIR mHealth and uHealth*, 10(5):e30517, May 2022. ISSN 2291-5222. doi: 10.2196/30517. URL https://mhealth.jmir.org/2022/5/e30517.

[7] Ricardo Ribeiro, Luísa Amaral, Wei Ye, Alina Trifan, António J. R. Neves, and Pedro Iglésias. Memoria: A memory enhancement and moment retrieval application for lsc 2023. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge*, LSC '23, page 18–23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701887. doi: 10.1145/3592573.3593099. URL https://doi.org/10.1145/3592573.3593099.

[8] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.00721. URL https://ieeexplore.ieee.org/document/10204762/.

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, June 2018. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2017.2723009. URL https://ieeexplore.ieee.org/document/7968387/.