FIRST - Flexible Interactive Retrieval SysTem for Visual Lifelog Exploration at LSC 2020

Minh-Triet Tran[1,3,4], Thanh-An Nguyen[1,4], Quoc-Cuong Tran[1,4], Mai-Khiem Tran[1,4], Khanh Nguyen[1,4], Van-Tu Ninh[2], Tu-Khiem Le[2], Hoang-Phuc Trang-Trung [1,4], Hoang-Anh Le [1,4], Hai-Dang Nguyen[1,4], Trong-Le Do[1,4], Viet-Khoa Vo-Ho[1,4], Cathal Gurrin[2]

[1]University of Science, Ho Chi Minh City, Vietnam
[2]Dublin City University, Ireland
[3]John von Neumann Institute, Ho Chi Minh City, Vietnam
[4]Vietnam National University, Ho Chi Minh City, Vietnam

ABSTRACT

Lifelog can provide useful insights of our daily activities. It is essential to provide a flexible way for users to retrieve certain events or moments of interest, corresponding to a wide variation of query types. This motivates us to develop FIRST, a Flexible Interactive Retrieval SysTem, to help users to combine or integrate various query components in a flexible manner to handle different query scenarios, such as visual clustering data based on color histogram, visual similarity, GPS location, or scene attributes. We also employ personalized concept detection and image captioning to enhance image understanding from visual lifelog data, and develop an autoencoder-like approach for query text and image feature mapping. Furthermore, we refine the user interface of the retrieval system to better assist users in query expansion and verifying sequential events in a flexible temporal resolution to control the navigation speed through sequences of images.

CCS CONCEPTS

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

KEYWORDS

lifelog; interactive retrieval; information system; component integration

ACM Reference Format:
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,and Hai-Dang Nguyen, Trong-Le Do, Viet-Khoa Vo-Ho, Cathal Gurrin. 2020. FIRST - Flexible Interactive Retrieval SysTem for Visual Lifelog Exploration at LSC 2020. In *Proceedings of the Twenty-Fifth International Conference on Architectural Proceedings of the Third Annual Workshop on the Lifelog Search Challenge (LSC '20), June 9, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3379172.3391726

1 INTRODUCTION

Lifelog [5] data provides valuable information for people to analyze their daily activities, events, and memories. However, it is not easy to describe what we want to search for from a massive amount of lifelog data, especially when much of the data is in a visual format. We can quickly enter a query in text format to retrieve text documents, but it is still a challenging problem to handle various types of queries, usually in text format, to look for a specific moment of interest in a collection of images or video clips.

The annual Lifelog Search Challenge (LSC [5]) aims to evaluate different approaches to assist users seeking to find events/moments of interest in an interactive manner from lifelog data. Many systems have been developed with different methods of image analysis and understanding, and different modalities for expressing queries and user interaction for result refinement [6].

Because of the wide variation of query types, such as location or scene-based queries, temporal sequence queries, or concept-based queries, the retrieval system should be flexible enough to add more query processing strategies or pipelines, or even support users to define their customized searching workflows. This motivates our proposal for a Flexible Interactive Retrieval SysTem (FIRST) to support different types of queries and to be open for future extensions.

Our proposed system is an enhancement to an existing system previously developed for the LSC2019 [10] with the following new features. Firstly, we refactor our legacy system and develop an integration platform that can be used to define and execute new query workflows and visualization layouts. For example, we can now visualize images into clusters with different semantic criteria, such as color histograms, scene attributes, or extracted deep features. Secondly, we improve our methods for image and scene understanding. We leverage our personalization strategy to generate captions for images by adapting our image captioning module to the personal lifelog data. We also propose an autoencoder-like method for mapping the features of a text query and an image to a common space to measure its semantic relationship. Thirdly, besides the default legacy query layout[10] , we create different visual layouts with clusters of images to assist users in searching for pictures in groups and add more interactions for users, such as query expansion by positive and negative examples. We also refine the user interface in [10] to better assist users in searching and verifying sequential events with flexible temporal resolution.

LSC '20, June 9, 2020, Dublin, Ireland
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7136-0/20/06…$15.00
https://doi.org/10.1145/3379172.3391726

The content of this paper is organized as follows. In Section 2, we briefly review related approaches for lifelogging retrieval. We introduce an overview of our system in Section 3 with the key idea to create a flexible integration platform to define and execute different pipelines. We introduce main methods to extract information from images in Section 4, including personalized concept detection and caption generation, as well as scene text extraction. The query layouts and interactions are presented in Section 5. The conclusion and discussion for future work are in Section 6.

2 RELATED WORK

Lifelog analysis and retrieval has become an attractive research topic in recent years. The first goal is to propose better methods for image/video understanding, and a second one is to develop more convenient modalities for users to interact with query systems. Challenges in different formats have been organized, targeting these two goals. In the recent ImageCLEF lifelog task [1] and the recent NTCIR14-Lifelog task [4], the retrieval tools are used by their creators who have in-depth knowledge about their systems, and the objective for participants mainly focuses on how to better extract information from visual lifelog data. In Lifelog Search Challenge (LSC [5]), the goals are not only to improve image understanding but also to design and enhance usability for professional and novice users to interact with the retrieval systems to handle various query types.

Successful systems have been developed and used in the Lifelog Search Challenge and the Visual Browser Showdown (VBS) [20]. The vitrivr system [19] supports different media types, such as images, videos, and audios, and also groups image sequences into segments for better representation. VIRET system [16] utilizes both visual data and non-visual information, such as weekdays, biometric data, GPS locations, to assist users in filtering certain events of interest. The SOM Hunter system [9] employs the technique proposed by Xirong Li[14] to compare the semantic distance between a text query and a video clip, which can be adapted for text query and image distance evaluation. Image clustering are also used in VIREO [18] and SOM-Hunter [9] systems. The system of the HCMUS team at LSC 2019 [12] enhanced the metadata by defining personalized visual concepts and creates an interface to navigate images in a sequence for temporal event verification. The Lifeseeker system [13] employed the Bag-of-Words model for text retrieval and query expansion using a Word2Vec model [17] pre-trained on GoogleNews dataset.

3 OVERVIEW OF FLEXIBLE INTERACTIVE RETRIEVAL SYSTEM

3.1 System Overview

Figure 1 illustrates the overview of our retrieval system. The essential feature in the system is the Flexible Integration Platform (see Section 3.2), which can be used to (1) define different component interaction as pipelines; (2) execute the workflow defined in a pipeline; and (3) convert and bridge data between components. This platform serves the other three main subsystems: Layout Manager, Query Component Manager, and MetaData Manager.

The MetaData Manager subsystem uses both visual and non-visual information, such as GPS data. For visual information, besides the concepts provided by the organizers, such as concepts extracted from Mask-RCNN [7] or scene attributes extracted using Place API, we extract concepts using CenterNet [3] and create a pipeline to train customized object detectors (see Section 4.1). We also refine the image captioning method [23] to adapt to the lifelog data.

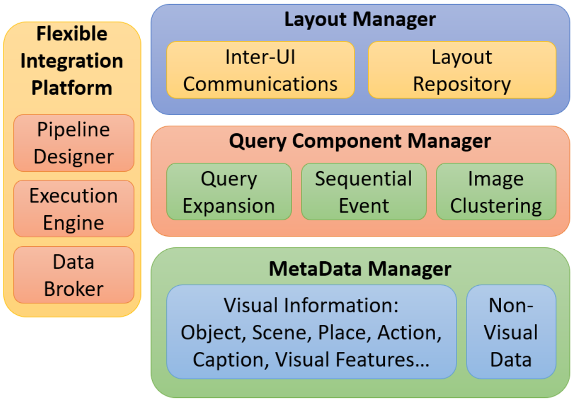


Figure 1: System overview of the query system

.

The Query Component Manager subsystem is responsible for processing different query types. Users can use the query expansion module to find and select concepts related to the query based on word embedding distance. The sequential event module exploits the temporal event relationship and frequent sequential events. The image clustering module helps users in interacting with visual data with different semantic distances, such as color histograms, deep visual features, or GPS locations, etc. Besides, in this subsystem, we can add more future-defined query processing pipelines using the Flexible Integration Platform.

In the Layout Manager subsystem, we create a repository of different layouts, including the traditional query interface [10], image cluster visualization layouts, etc. We also create a communication mechanism between different UI components to allow flexible inter-UI interaction.

3.2 Flexible Integration Platform

By using an integration platform, we can customize the process to handle various query types. Users can select the workflows or pipelines that they think the most suitable for the problem by choosing available components and define how these components connect to others. Users can communicate to the integration platform efficiently by dragging and dropping interaction, then execute the defined pipeline to process queries.

The platform consists of a diagram library for an end-user to quickly create a runnable workflow with minimal effort and to monitor the process when running. On the back-end side, we have an engine for managing the workflow, including component integration, data flow control, logging.

Each component has to support to notify the current state: initializing, ready, executing, finished. When a component finishes, its next components in the pipeline receive the output as their inputs to begin their execution.

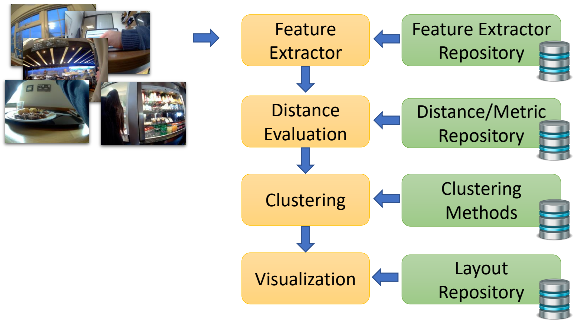


Figure 2: Pipeline for visual/semantic distance estimation and visualization.

We define multiple atomic components for different tasks: feature extractors, image clustering, visualizer, etc. Components that support the same mission have the same format for input and output. Workflow data come from various multiple forms, such as images, videos, vector, etc. We also pre-define several common best practice flows from our experience that can be selected for use by users.

Our platform also supports the workflow with parallel branches by using if-condition. For example, we can check if the location vector has 2-dimension, then we can use a grid visualizer, but if it has 3-dimension or more, we can use VR for visualization.

3.3 Flexible Image Clustering

In [22], we proposed to group images based on BoW- similarity. In this way, we can have clusters of images, and each may correspond to a working place. From this strategy, we continue to develop a more flexible image clustering technique. Our extension is also inspired by image clustering functions in VIREO [18] and SOM-Hunter [9] systems.

Figure 2 illustrates our pipeline to create different strategies to estimate the visual/semantic distance between images. By using our integration system, we can easily change and test with other feature extractors such as EfficientNet [21] or ResNet [8]. We also can test will multiple dimension reduction algorithms, such as PCA. For example, we use ResNet50 with weights trained on ImageNet to extract features and multi-dimensional scaling with pairwise Euclidean distance for dissimilarity measure to reduce the dimension to 2D.

4 INFORMATION EXTRACTION FROM IMAGES

4.1 Personalized Concept Detection

In our previous work [10, 11], we proposed to detect and extract personalized visual concepts from visual lifelog data. We first manually extract personalized concepts of a lifelogger, i.e., his or her everyday visual objects that are not available in public visual datasets, and train object detectors using Faster-RCNN. In this way, our system can better adapt to the personal life of a lifelogger and can detect such visual concepts from his or her lifelog data.

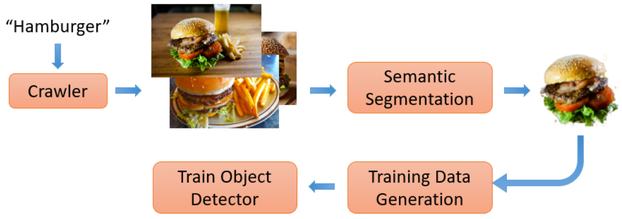


Figure 3: Pipeline for creating personalized object detector.

In this paper, we take further steps to enhance the personalization for lifelog analysis. Firstly, we create a pipeline to automatically crawl images from the Internet which are corresponding to a new visual concept, such as čoffee machineör ȟamburger, extract the main content using semantic segmentation, embed objects into different backgrounds, train object detectors, and apply the detectors on lifelog dataset. In this way, we can boost up the process to define numerous personalized object detectors adapting to the lifelogger's habits (see Figure 3).

4.2 Personalized Caption Generation

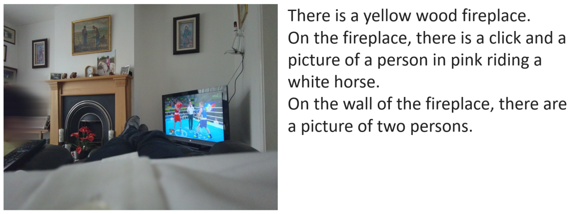
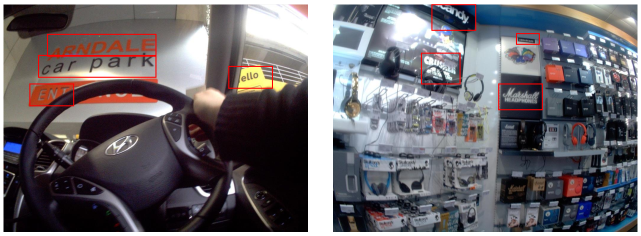

There is a yellow wood fireplace.
On the fireplace, there is a click and a picture of a person in pink riding a white horse.
On the wall of the fireplace, there are a picture of two persons.

Figure 4: An example of a personalized caption.

We utilize our proposed method for concept-augmented image captioning [23] to train a captioning module based on a small dataset of captioning for a lifelogger. We randomly select a subset of 1,000 images from the lifelog dataset and manually annotate their captions. We have 2-3 volunteers annotating each image. In this way, we create a small but sufficient volume of data to refine our captioning module. An example of generated personalized captions is shown in Figure 4. Then we use our refined captioning module to generate captions for the whole lifelog dataset. The dense captioning strategy is also used to create more descriptions for different regions in an image. The generated captions are stored in our database to be matched against a future query text or phrase.

4.3 Scene Text Extraction

To better understand information from an image, we use Convolutional Character Networks [24] to extract scene texts from an image, such as brand names of products in a shop, shop names, street names, etc. We also use ABCNet [15] to handle texts in Bezier-Curve shapes. Figure 5 shows two examples of extracted texts from images.



arndale, car park, ent, ello

candy, marshall, headphones, crus, urbanears

Figure 5: An example of a personalized caption.

As recognized text can be occluded, we use a dictionary to autocorrect and the edit distance to compare the similarity between a query text and extracted text from an image. Furthermore, we employ Word2Vec to evaluate the semantic relationship between words/phrases.

4.4 Mapping Query Text and Images to An Invertible Common Feature Space with AutoEncoder

To compare the similarity between a query text and an image, we can encode the query and the image into the sentence and image features, respectively. We can use BERT for sentence embedding, and ResNet for image embedding. A common approach is to map these two features into a common space to measure the distance between them, then infer the relationship between the query text and the image [14].

Figure 6 shows the overview of our proposed method to map a query text of an image to a common feature space with autoencoder. An important property of the common space is that it can be used to measure the distance or dissimilarity between the features of two different data types (or domains).

In our proposed method, we aim to achieve one more property for the common space: the information of a feature in the common space should be enough to approximately reconstruct the origin feature, either from a text query or an image. Therefore, together with two mapping functions from the sentence embedding and the image embedding to the common space, we also define the two inverse mapping functions from the common space back to the feature spaces for sentences and images.

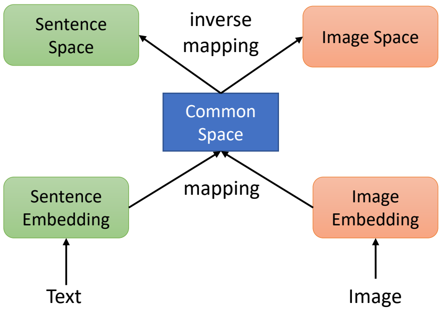


Figure 6: AutoEncoder-like approach for mapping query text and image to a common feature space

.

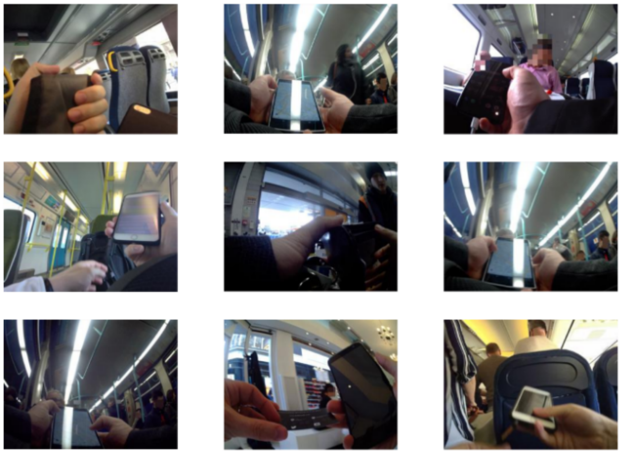


Figure 7: Retrieved images for "using cellphone in a train".

Our auto-encoder like approach is inspired from the idea of the dual encoding for zero-example video retrieval [2]. In Figure 7, we demonstrate the result of the query to search for "using cellphone in a train" using our proposed method for text and image mappings with autoencoder.

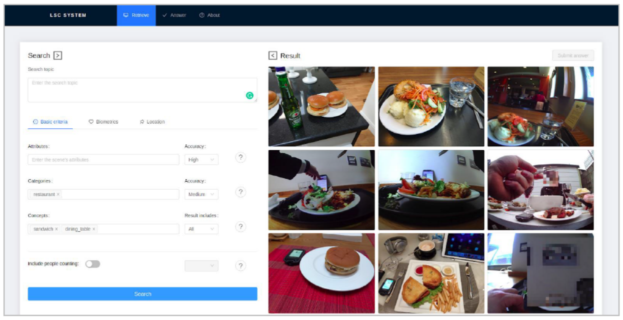
5 QUERY LAYOUT AND INTERACTION

5.1 Query Layouts

The presentation subsystem plays an essential part in visualizing the results intuitively because a clear representation of data can have a significant impact on the insights derived from queries.

With expandability and usability in mind, we build the presentation subsystem as a collection of modules working in tandem, processing, and rendering queries from the source to the results. As such, it covers several configurable parameters.

While the sheer number of parameters provides flexibility over the system, it also has the potential to worsen the usability due to its complexity. To avoid such problems, we have several presets which are preloaded that suit basic visualizing scenarios.



Search topic: Enter the search topic

Attributes: Enter the scene's attributes — Accuracy: High

Categories: restaurant × — Accuracy: Medium

Concepts: sandwich × dining_table × — Result includes: All

Include people counting:

Figure 8: Legacy layout as the default mode [11]

.

Our current system is developed from our previous retrieval system [10]. For the LSC in 2020, we reuse our query interface in [11], as shown in Figure 8. We keep the default layout because of its simplicity for users to use, but we add more interaction functions into the system.

In the default mode, users can input keywords or phrases to query for a moment. Together with new techniques to extract more concepts from images (see Section 4), we also adopt the semantic similarity between a query phrase and an image. Our method is inspired by the work of Xirong Li et al. [14]. The key idea is to encode a query phrase as a feature in the feature space of images. In this way, we can measure the distance between the encoded feature of a query phrase and the visual feature of an image. We also propose a new method for this function in Section 4.4.

5.2 Image Cluster Visualization

In Figure 9, we illustrate the layout in our system to visualize images based on their semantic similarity. In our system, the semantic similarity can be determined in different ways, such as GPS, Bag-of-Visual-Word features, scene attributes, or color histograms.

We also support grouping a sequence of consecutive images into an event, an allow users to change the level of details in the image cluster visualization layout. In the global view, images are grouped in several main clusters. When a user selects an image cluster, it is expanded to show its subgroups. A user can explore image clusters in a coarse-to-fine approach.

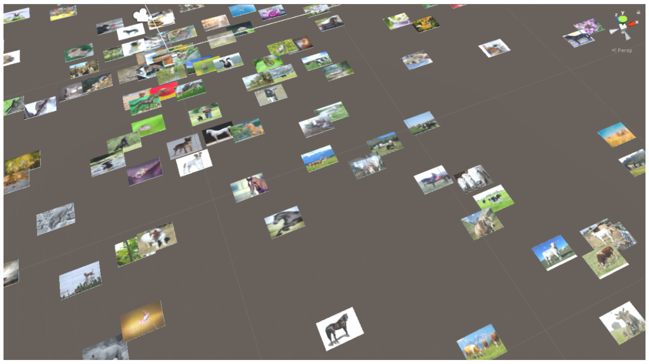


Figure 9: Image Visualization based on Semantic Similarity.

5.3 Example-based Query Expansion

It would be a convenient way to assist users in finding moments in a lifelog data having similar content with a sample image. For example, when we find an example image of watching TV at home, we can use this image as a positive example of finding all similar moments.

In this way, we can replace multiple query criteria to find the example image by that image to further retrieve other images or moments. We also allow users to select an image as a negative example to eliminate images that are similar to it.

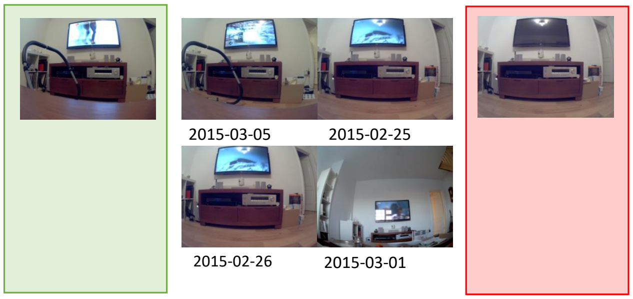


Figure 10: Example-based query expansion to find images/moments similar or different from examples.

For any image displayed in the user interface of our system, we provide a query expansion mechanism from that image. A selected image is considered as a positive or negative example for query expansion. As illustrated in Figure 10, we want to find the moments of watching TV, and the TV must be turned on. In the retrieved list, images or moments are retrieved and sorted in descending order of similarity with positive examples (in the green box: TV is on), then in ascending order of dissimilarity with negative examples (in the red box: TV is off).

5.4 Flexible Temporal Resolution for Sequence of Events Exploration

In [10], we created a user interface to assist users in surfing a sequence of images (backward and forward) from any given moment in the lifelog data to look for certain other events before or after the selected moment. However, in this implementation, we notice that it is time-consuming for a user to navigate to a long-distance event as we use a fixed time step. Thus, in the current system, we allow users to easily adjust their navigation step by using a temporal step slider. We expect this way can help users in controlling the operation that best matches their needs.

6 CONCLUSION AND FUTURE WORK

In this paper, we introduced our flexible system for lifelog data retrieval. The key idea of our system is to enhance the flexibility to define different pipelines to handle both visual data understanding and query processing.

We use the designer in Flexible Integration Platform to define then execute various workflows that might be useful for a wide variation of query types. We also leverage the personalization in visual lifelog data analysis with adapted image captioning, and utilize the text and visual distance comparison [14] to retrieve images with captions semantically related to a text query.

Currently, based on our subjective experience, we predefined several pipelines to illustrate the usability of the Flexible Integration Pipeline in image analysis and query processing. We expect that the flexible mechanism of our system can efficiently assist researchers and users in developing and integrating their own components into the platform and defining then executing their customized workflows for multimedia analysis and retrieval.

ACKNOWLEDGMENTS

This research is partially supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19 and Vietnam Ireland Bilateral Education Exchange Programme (VIBE) 2019. We would like to thank AIOZ Pte Ltd for supporting our research team with computing infrastructure.

REFERENCES

[1] Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Liting Zhou, Mathias Lux, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. 2019. Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In *CLEF2019 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland.

[2] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2018. Dual Encoding for Zero-Example Video Retrieval. arXiv:cs.CV/1809.06181

[3] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. CenterNet: Keypoint Triplets for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 6568–6577. https://doi.org/10.1109/ICCV.2019.00667

[4] Cathal Gurrin, H. Joho, Frank Hopfgartner, Liting Zhou, Tu Ninh, Tu-Khiem Le, Rami Albatal, D.-T Dang-Nguyen, and Graham Healy. 2019. Overview of the NTCIR-14 Lifelog-3 task.

[5] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoč, Wolfgang Hurst, Minh-Triet Tran, and Klaus Schoeffmann. 2020. An Introduction to the Third Annual Lifelog Search Challenge, LSC'20. In *ICMR '20, The 2020 International Conference on Multimedia Retrieval*. ACM, Dublin, Ireland.

[6] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Liting Zhou, Aaron Duane, Andreas Leibetseder, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[8] 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

[9] Miroslav Kratochvíl, Patrik Veselý, Frantisek Mejzlík, and Jakub Lokoc. 2020. SOM-Hunter: Video Browsing with Relevance-to-SOM Feedback Loop. In *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II (Lecture Notes in Computer Science)*, Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve (Eds.), Vol. 11962. Springer, 790–795. https://doi.org/10.1007/978-3-030-37734-2_71

[10] Nguyen-Khang Le, Dieu-Hien Nguyen, Trung-Hieu Hoang, Thanh-An Nguyen, Thanh-Dat Truong, Tung Dinh Duy, Quoc-An Luong, Viet-Khoa Vo-Ho, Vinh-Tiep Nguyen, and Minh-Triet Tran. 2019. Smart Lifelog Retrieval System with Habit-based Concepts and Moment Visualization. In *Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2019, Ottawa, ON, Canada, 10 June 2019*, Cathal Gurrin, Klaus Schöffmann, Hideo Joho, Duc-Tien Dang-Nguyen, Michael Riegler, and Luca Piras (Eds.). ACM, 1–6.

[11] Nguyen-Khang Le, Dieu-Hien Nguyen, Vinh-Tiep Nguyen, and Minh-Triet Tran. 2019. Lifelog Moment Retrieval with Advanced Semantic Extraction and Flexible Moment Visualization for Exploration. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019 (CEUR Workshop Proceedings)*, Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller (Eds.), Vol. 2380. CEUR-WS.org. http://ceur-ws.org/Vol-2380/paper_139.pdf

[12] Nguyen-Khang Le, Dieu-Hien Nguyen, Trung-Hieu Hoang, Thanh-An Nguyen, Thanh-Dat Truong, Duy-Tung Dinh, Quoc-An Luong, Viet-Khoa Vo-Ho, Vinh-Tiep Nguyen, and Minh-Triet Tran. 2019. Smart Lifelog Retrieval System with Habit-Based Concepts and Moment Visualization. Association for Computing Machinery, New York, NY, USA.

[13] Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Minh-Triet Tran, Liting Zhou, Pablo Redondo, Sinéad Smyth, and Cathal Gurrin. 2019. LifeSeeker: Interactive Lifelog Search Engine at LSC 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC@ICMR 2019, Ottawa, ON, Canada, 10 June 2019*, Cathal Gurrin, Klaus Schöffmann, Hideo Joho, Duc-Tien Dang-Nguyen, Michael Riegler, and Luca Piras (Eds.). ACM, 37–40. https://doi.org/10.1145/3326460.3329162

[14] 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, MM 2019, Nice, France, October 21-25, 2019*, Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi (Eds.). ACM, 1786–1794.

[15] Yuliang* Liu, Hao* Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. 2020. ABCNet: Real-time Scene Text Spotting with Adaptive Bezier-Curve Network. In *Accepted to Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020*.

[16] Jakub Lokoč, Tomáš Souček, Premysl Čech, and Gregor Kovalčík. 2019. Enhanced VIRET Tool for Lifelog Data. Association for Computing Machinery, New York, NY, USA.

[17] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. 1–12.

[18] Phuong Anh Nguyen, Jiaxin Wu, Chong-Wah Ngo, Danny Francis, and Benoit Huet. 2020. VIREO @ Video Browser Showdown 2020. In *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part II (Lecture Notes in Computer Science)*, Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve (Eds.), Vol. 11962. Springer, 772–777. https://doi.org/10.1007/978-3-030-37734-2_68

[19] Luca Rossetto, Ralph Gasser, Silvan Heller, Mahnaz Amiri Parian, and Heiko Schuldt. 2019. Retrieval of Structured and Unstructured Data with Vitrivr. Association for Computing Machinery, New York, NY, USA.

[20] K. Schoeffmann. 2019. Video Browser Showdown 2012-2019: A Review. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–4. https://doi.org/10.1109/CBMI.2019.8877397

[21] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 6105–6114. http://proceedings.mlr.press/v97/tan19a.html

[22] Thanh-Dat Truong, Tung Dinh Duy, Vinh-Tiep Nguyen, and Minh-Triet Tran. 2018. Lifelogging Retrieval based on Semantic Concepts Fusion. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, LSC@ICMR 2018, Yokohama, Japan, June 11, 2018*, Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Duc-Tien Dang-Nguyen, Michael Riegler, and Luca Piras (Eds.). ACM, 24–29. https://doi.org/10.1145/3210539.3210545

[23] Viet-Khoa Vo-Ho, Quoc-An Luong, Duy-Tam Nguyen, Mai-Khiem Tran, and Minh-Triet Tran. 2018. Personal Diary Generation from Wearable Cameras with Concept Augmented Image Captioning and Wide Trail Strategy. In *Proceedings of the Ninth International Symposium on Information and Communication Technology, SoICT 2018, Danang City, Vietnam, December 06-07, 2018*. ACM, 367–374.

[24] Linjie Xing, Zhi Tian, Weilin Huang, and Matthew R Scott. 2019. Convolutional Character Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.