UA.PT Bioinformatics at ImageCLEF 2020: Lifelog Moment Retrieval Web based Tool

Ricardo Ribeiro, Júlio Silva, Alina Trifan, José Luis Oliveira, and António J. R. Neves

IEETA/DETI, University of Aveiro, 3810-193 Aveiro, Portugal
{rfribeiro, silva.julio, alina.trifan, jlo, an}@ua.pt

Abstract. This paper describes the participation of the Bioinformatics group of the Institute of Electronics and Engineering Informatics of University of Aveiro in the ImageCLEF lifelog task, more specifically in the Lifelog Moment Retrieval (LMRT) sub-task. In our first participation last year we tackled the LMRT challenge with an automatic approach. Following the same steps, we improved our results, while introducing a new interactive approach. For the automatic approach, two submissions were made. We started by processing all images in the lifelog dataset using object detection and scene recognition algorithms. Afterwards, we processed the query topics with Natural Language Processing (NLP) algorithms in order to extract relevant words related to the desired moment. Finally, we compared the visual concepts of the image with the textual concepts of the query topic with the goal of computing a confidence score that relates the image to the topic. For the interactive approach, we developed a web application in order to visualize and provide an interactive tool to the users. The application is divided in three stages. In the first one, the user uploads the images from the dataset, as well the textual data annotations. In the second stage, the user interacts with the application assigning the extracted words to the several topics. Consequently, the application retrieves the image associated to the topic with a certain confidence. In the last stage, we provide a visual environment with two different views, in the form of a image gallery or data tables organized into timestamp clusters. Similarly to our previous participation, the results of the automatic approach are still far from being competitive. We conclude that an automatic approach might not be the best solution for the LMRT task since the currently available state-of-the-art technology is still not able to wield better results. However, our interactive approach with relevance feedback obtained better and competitive results, achieving a F1-measure@10 score of 0.52.

Keywords: lifelog · moment retrieval · image processing · web application

1 Introduction

The number of workshops and tasks for research has increased over the last few years and among them are the main fields of ImageCLEF 2020 lab [3]: multimedia retrieval in lifelogging, medical, mature, and internet applications. The multimedia retrieval in lifelogging has received significant attention from both research and commercial communities. The increasing number of mobile and wearable devices is dramatically changing the way we collect data about a person's life.

Lifelogging is defined as a form of pervasive computing consisting of a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive. In a simple way, lifelogging is the process of tracking and recording personal data created through our activities and behaviour [1].

Personal lifelogs have a great potential in numerous applications, including memory and moments retrieval, daily living understanding, diet monitoring, or disease diagnosis, as well as other emerging application areas [9]. For example: in Alzheimer's disease, people with memory problems can use a lifelog application to help a specialist follow the progress of the disease, or to remember certain moments from the last days or months.

One of the greatest challenges of lifelog applications is the large amount of lifelog data that a person can generate. The lifelog datasets, for example the ImageCLEFlifelog dataset [5], are rich multimodal datasets which consist in one or more months of data from multiple lifeloggers. Therefore, an important aspect is the lifelog data organization in the interest of improving the search and retrieval of information. In order to organize the lifelog data, useful information has to be extracted from it. Other important aspects are the visualization and user interface of the application.

With the purpose of improving the results obtained in the previous year's challenge [7], we developed a first version of a web application to provide a visual and interactive environment to the user. In last year's work [7], the approach was fully automatic using an exhaustive method to retrieve data and there was no tool for visualization and interaction with the user. However this year, a significant improvement has been made with regard to the data retrieval using a dynamic and faster method. Initially, only the data provided by the organization is used and stored in the database to further use in the retrieval stage in our application. We divided this approach into 3 different stages, such as upload, retrieval and visualization. At each stage, there is an interaction with the user, which is encouraged by the organizers of the ImageCLEFlifelog [5]. The web application is still in an early stage but is the baseline of our current work.

This paper starts with an introductory section and it is organized as follows. Section 2 provides a brief introduction to the ImageCLEF lifelog and the subtask Lifelog Moment Retrieval. The proposed methods are described in Section 3. In Section 4, the results of all submitted runs obtained in the LMRT sub-task are described. Finally, a summary of the work presented in this paper, concluding remarks, and future work can be read in Section 5.

2 Task Description

The ImageCLEFlifelog 2020 task [5] is divided into two different sub-tasks: the Lifelog moment retrieval (LMRT) and Sport Performance Lifelog (SPLL) subtask. In this work, as in the previous year's challenge [7], we only addressed the LMRT sub-task, as a continuous research work that we intend to develop with the aim of giving our contribution to real problems that exist around the world that can benefit from this technology.

In the LMRT subtask, the main objective is to create a system capable of retrieving a number of predefined moments in a lifelogger's day-to-day life from a set of images. Moments can be defined as semantic events or activities that happen at any given time during the day. For example, given the query "Find the moment(s) when the lifelogger was having an icecream on the beach" the participants should return the corresponding relevant images that show the moments of the lifelogger having icecream at the beach. Like last year, particular attention should be paid to the diversification of the selected moments with respect to the target scenario.

ImageCLEFlifelog dataset is a new rich multimodal dataset which consists of 4.5 months of data from three lifeloggers, namely: images (1,500-2,500 per day), visual concepts (automatically extracted visual concepts with varying rates of accuracy), semantic content (locations and activities) based on sensor readings on mobile devices (via the Moves App), biometrics information (heart rate, galvanic skin response, calories burn, steps, continual blood glucose, etc.), music listening history and computer usage [5]. However, in this work we only use the images, the visual concepts and the semantic content of the dataset.

3 Proposed Method

We submitted a total of 3 runs in the LMRT sub-task. The work made this year had a significant improvement comparing with our previous work [7], due to the interactive and visual approach with the user that we choose to apply. In this section, we present the proposed approach of our submissions. The first two runs follow the same approach as last year [7], where we aimed at building a fully automatic process for image retrieval. However, the improvement is in our last submission, in which a web application was developed providing visual and interactive environment to the user. This web application is a first prototype, far from a final version, but we consider it as a baseline of our work.

3.1 Automatic approach (Run 1 and 2)

Initially, the images of the dataset were processed using algorithms for label detection, such as objects and scenes. The information provided by the organizers, such as locations, activities and local time, are also used. In both runs, for scene recognition we used a pretrained model provided by Zhou et al. [10] trained on the Places365 standard dataset. For the first run, the method used to extract objects from the images is a combination of ResNeXt-101 and Feature Pyramid Network architectures in a basic Faster Region-based Convolutional Network (Faster R-CNN) pretrained on the COCO dataset that was proposed by Mahajan et al. [4].

In the second run, the object detection algorithm used is the YoloV3 [6] model pretrained in the COCO dataset. Subsequently, we proceed to the extraction of relevant words from the query topics and the computation of the semantic similarity between word vectors done with a Natural Language Processing library called SpaCy [2]. From the topic title, description and narrative, relevant words were extracted and organized into different categories, such as relevant things, negative things, activities, dates, locations and environment.

Using topic 1 as an example :

- **Title** : ”Praying Rite.”
- **Description** : ”Find the moment when u1 was attending a praying rite with other people in the church.”
- **Narrative** : ”To be relevant,the moment must show u1 is currently inside the church, attending a praying rite with other people. The moments that u1 is outside with the church visible or inside the church but is not attending the praying rite are not considered relevant.”

The extracted textual data is as follows:

- relevant things - ”rite” , people”.
- activities - ”praying”, ”praying rite”, ”attending”.
- locations - ”church”
- dates - empty.
- user inside - ”true”.
- user outside - ”false”.
- negative relevant thing - ”church visible”.
- negative locations: empty.
- negative activities : empty.
- negative dates: empty.

Afterwards, a confidence score is computed for each image in the dataset. The score is obtained through the comparison of the extracted words from the topic and the extracted labels from the images. This score is influenced by the scores of the image concepts obtained through the object detection phase and the different weights assigned to each category. The weight for each category is obtained through two different factors, a factor of importance and a computed factor.

In Run 1, the importance factor for all categories is the same. This means that each category has the same weight for the computation of the confidence score.

For Run 2, we decided to define the importance factor differently for each category. We give a bigger importance to specific categories like ”relevant things” in order to improve results, since we compute the similarity of this textual category with our object detection extracted image label concepts. Categories like "activities" and "locations" get a lesser importance factor since they are being compared to the organizers label data which is limiting and lesser accurate. The sum of all importance factors of all categories is equal to 1, which represents 100%.

The computed factor is obtained from the distribution of the factor of importance from empty categories to all other categories. If we don't extract any textual data from a query topic for the category "activities", this category will be empty, therefore, we apportioned the importance factor of the "activities" category to all other categories, increasing their importance factor, in order to maintain the sum of 1. This value is not the same for each category, we maintain the ratio of the distribution the same as the distribution of the importance factor between all categories. To make it clearly, if the importance factor for "relevant things" is 0.5, which is half of the sum of all importance factors, and if the "activities" category is worth 0.2 and has no extracted textual data, then half of 0.2 is distributed to "relevant things", which increases the importance to 0.6 and the remainder 0.1 will be distributed the same way to other categories ensuring that the sum of all importance factors is 1.

The negative categories works the same way, but instead of contributing for the confidence score, it decreases the value of the confidence.

A general threshold was previously defined in order to remove images of low concept scores or low confidence score, images above the threshold are selected for the query topic. The threshold was implemented through some trial and error during the test phases, and it merely serves the purpose of saving some computational time.

Run 2 differs from Run 1 not only in the image processing step, where different image processing algorithms were used, but also in the retrieval step, where all factors of importance were altered in order to give more importance to some categories than others, as previously discussed. Another difference is the negative category which was discarded from the calculation of the confidence score in Run 2.

Finally, a script runs through all the selected confidence scores for a given query topic and stores the fifty highest on the csv file. As expected by the previous year results, this automatic and exhaustive approach is not the most suitable for a lifelog application.

3.2 Web application (Run 3)

To improve our results in this challenge, we develop a web application in order to visualize and provide an interactive tool for our lifelog system. As encouraged by the organizers, in this run, we used a method that allows interaction with users. As a first approach, we are only considering the data provided by the challenge organizers. We divided the web application into three stages, respectively:

- **Upload:** the user uploads the images from the lifelog dataset into the application. The textual data annotations provided by the organizers are automatically uploaded and organized in the application database associated with uploaded images.

- **Retrieval:** the user introduces the inputs words extracted from the query topic into several words categories, date and time. The retrieval process starts comparing these inputs with the app database information. Finally, a confidence to each image retrieved is assigned for the query topic.
- **Visualization:** the user visualizes the retrieved images and scores, in form of image gallery or data tables, divided into timestamp clusters. The user choose manually the relevant clusters for the query topic.

Figure 1 shows a general representation of our lifelog application. In a first stage, the user has to upload the images into the application, which are stored in the database together with the data provided by the organizers for each image from the lifelog dataset. Afterwards, the user requests the image retrieval for the query topic by introducing relevant words manually in the application, the stage of retrieval begins. These relevant words are divided into several categories, such as objects, locations, activities, irrelevant words, date and time, and they are compared with the labels stored in the database. This comparison is made using the similarity of word vectors. Images with labels similar to the topic relevant words are selected. Subsequently, the confidence for the corresponding image is computed through the similarity value of the labels and the score of each similar label in the database. In order to reduce the amount of images retrieved by the system, images with low confidence are excluded from the output images. At the end, the retrieved images are clustered by timestamp intervals and the user can visualize the images in the form of image gallery or data tables.

A more detailed explanation is provided in the following sections for each stage of our lifelog application.

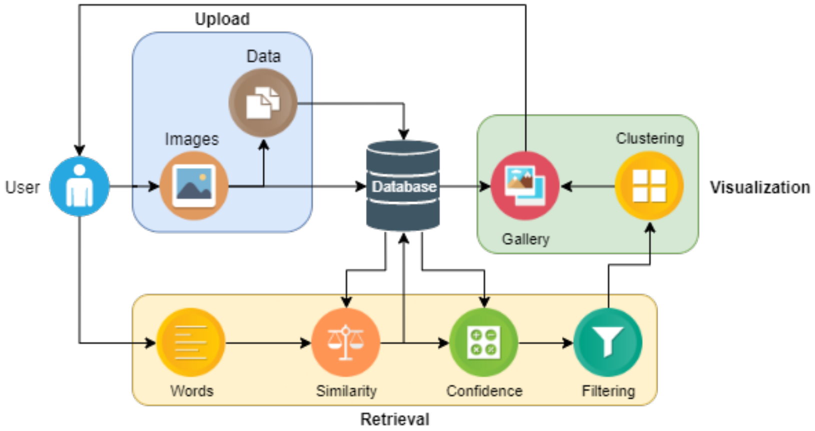


Fig. 1. General representation of the developed web application. The user interacts with the three stages of the application: Upload, Retrieval and Visualization.

Upload In an initial stage, the user uploads the images dataset into the lifelog application that are organized and stored in the database associated with some of the data provided by the organizers, such as visual concepts and metadata. The data is organized in our database into different tables/models, such as images, concepts, locations, activities, scenes, attributes, among others. In our application, each model maps to a single database table. Figure 2 shows a diagram of these data models in the database. The relationship between models makes our system faster and more efficient compared to an exhaustive approach.

The image model has a many-to-many relationship with the models concept, location, category, activity and attribute. For example: an image can contain several concepts, and a concept can be found in several images. The tag field of the label model is the labels name extracted from the visual concepts and metadata, which has a one-to-many relationship with the other models, in other words, one label may be connected to several images and this label can be associated to several models, such as concept and category models, depending on the type of label and the number of times that appear in the image. Usually, the name of the labels are in their base form or dictionary form, called the word's lemma, however labels in other forms are transformed to the basic form for further use. This transformation is called lemmatizer.

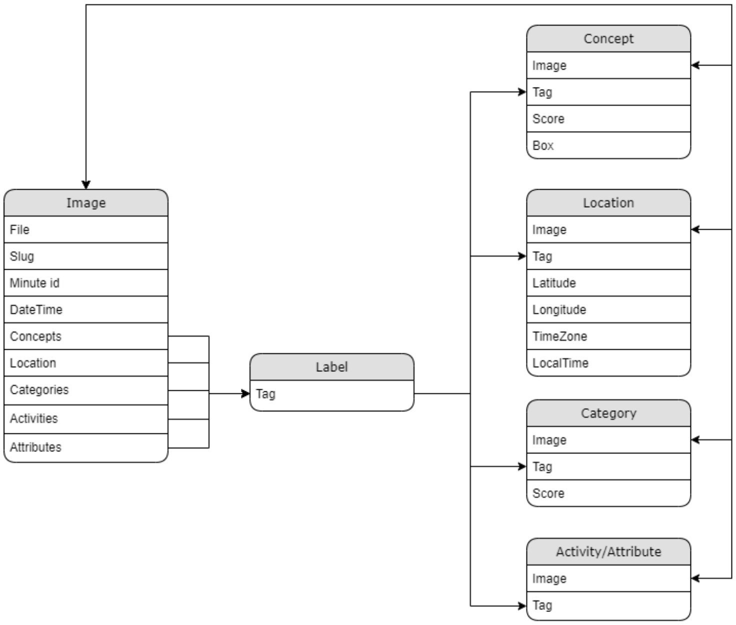


Fig. 2. Diagram of the proposed database tables used by the web application.

Retrieval Unlike the exhaustive approach of run 1 and 2, that compute the confidence of each image, this approach (run 3) only computes the confidence of some images that are selected in a first step for the specific topic by using the similarity of word vectors, which makes this retrieval method more efficient and using less processing time.

The topics are manually analysed by the user, which extracts relevant words from them. By introducing these words divided into several categories, such as objects, locations, activities and irrelevant words in the application, the retrieval step begins. If a topic contains time ranges, years or days of the week, the user can also insert that data in our application to further filter the retrieved images. Figure 3 shows the retrieval view of the web application.

In the retrieval stage, the input arguments are: objects that appear on the images; activities that the user was practicing; locations or places where the user was; negatives or irrelevant things, activities or locations that should not appear in the images; time ranges, years and days of the week (Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday).

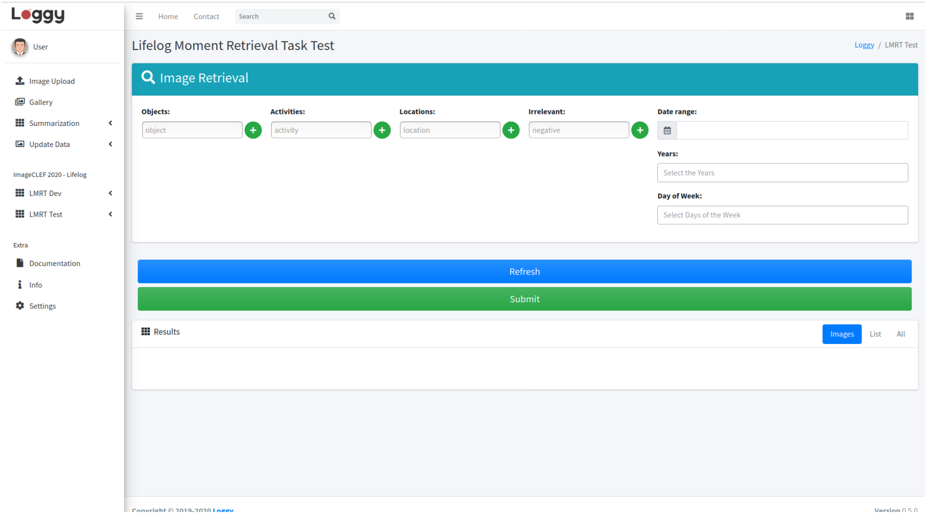


Fig. 3. Web application retrieval view.

The SpaCy library [2] is used for two different tasks: to assign the base forms words (lemmatizer) and to compare word vectors (cosine similarity). As in the upload stage, the input words are processed to their lemma, which improves and facilitate the comparison between word vectors. Afterward, the similarity between the processed input words and the labels stored the database is computed. Images that contains labels that are similar or equal to the words entered by the user are selected to compute the confidence. If the user enters negative words in our applications, images with labels similar or equal to these negative words are automatically excluded. In order to improve the processing time of the retrieval stage, the similarity of word pairs are stored in the database so that it is not necessary to compute the similarity of the same word pair more than once.

The confidence of the selected images is computed using the similarity calculated previously and the score of the labels. For labels without score field, it is only used the similarity to calculate the confidence. As last filtering on the retrieval stage, the images are selected based on the confidence threshold.

Visualization The selected images are organized into different clusters based on images timestamps provided by the organizers. The retrieved images were visualized in our application organized into the timestamp clusters. The application provides an easy way for users to visualize and identify the clusters that are associated to the specific topic. Figure 4 shows the user view of the clustered images in form of images gallery. We provided another way of visualization in form of a data table as shown in Figure 5.

In order to improve the results, the user can exclude several irrelevant images from the selected clusters. To improve the cluster recall of the run, the user can change the confidence of a relevant image of each selected timestamp clusters to the maximum confidence that consequently increases the f1 measure of this run.

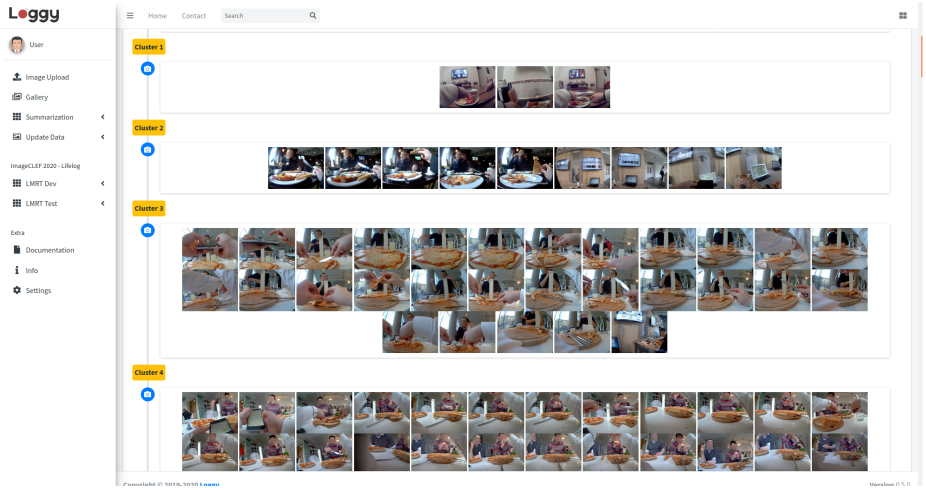
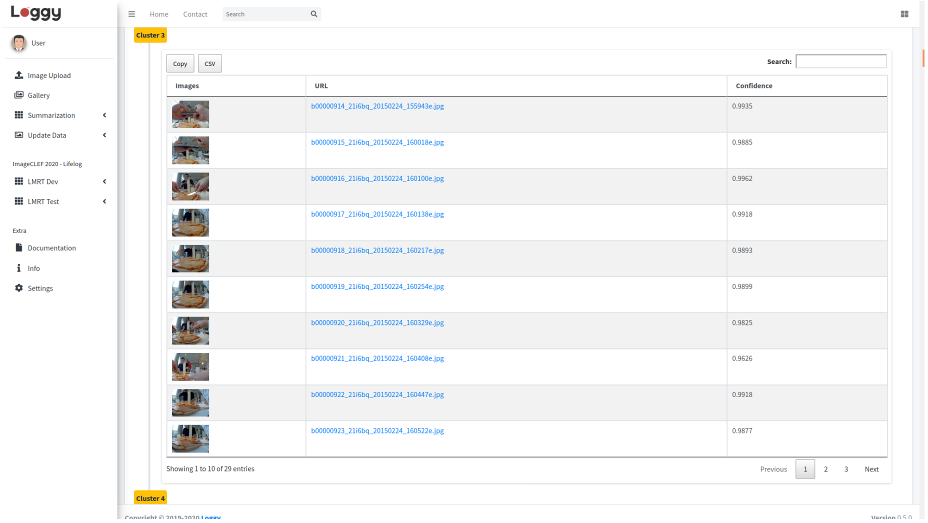


Fig. 4. User view of the image clusters in form of image galleries.



Images	URL	Confidence
	b00000914_21i6bq_20150224_155943e.jpg	0.9935
	b00000915_21i6bq_20150224_160018e.jpg	0.9885
	b00000916_21i6bq_20150224_160100e.jpg	0.9962
	b00000917_21i6bq_20150224_160138e.jpg	0.9918
	b00000918_21i6bq_20150224_160217e.jpg	0.9893
	b00000919_21i6bq_20150224_160254e.jpg	0.9899
	b00000920_21i6bq_20150224_160329e.jpg	0.9825
	b00000921_21i6bq_20150224_160408e.jpg	0.9626
	b00000922_21i6bq_20150224_160447e.jpg	0.9918
	b00000923_21i6bq_20150224_160522e.jpg	0.9877

Showing 1 to 10 of 29 entries

Fig. 5. User view of the image clusters in form of data tables.

4 Results

We submitted a total of 3 runs on the LMRT sub-task. In this task, an arithmetic mean of all query topics results is calculated as the final score. The ranking metrics was the F1-measure@10, which gives equal importance to diversity (via CR@10) and relevance (via P@10), Cluster Recall and Precision at top 10 results, respectively.

We described the three submissions in Section 3. The first two submissions follows an automatic manner as in our previous work [7]. Due to the results of this automatic approach, we take into consideration the development of a system that allows interaction with real users, as emphasized by the organizers.

Comparing the automatic with the interactive approach, a significant improvement can be seen. This improvement is due to not only to the new retrieval approach, but also to the interactive and visual approach. We consider the visualization and user interaction one of the most important tools in a lifelog application.

4.1 UA.PT Bioinformatics Results

The results obtained are shown in Table 1, along with the best result in this task, for comparison. The results of all of the participating teams can be found in [5]. We can observe that our last submission (run 3) is still not the best on this task, but we made a considerable improvement compared with the automatic approach from this year and the previous year [7], and we are also closing the gap between the best ones, such as HCMUS team with the best F1-measure@10 on the LMRT task, with the ambition to obtain much better results.

Table 1. F1-measure@10 of each run submitted by us and the best team run in the LMRT task.

Team	Run Name	F1-measure@10
Our	Run 1	0.03
	Run 2	0.03
	Run 3	0.52
HCMUS	Run 10	0.81

Considering the results shown in Table 1 we are convinced that the interactive approach is a better suited method for the LMRT challenge, the user visualization and interaction with the application allows for much more accurate results. Creating a fully automatic system is complicated, this is because it requires a lot of processing power, every image has to be fully processed in order to extract labels. However, considering that computing time is not a problem, a few ways that we could improve the results of our automatic approach in the future would be implementing activity recognition algorithms, color recognition algorithms and better scene recognition algorithms.

As an initial lifelog application, the results shows that we are in a good path to solve some of the problems that exist in these challenges, which could help to improve the daily lives of many people. Considering the previous work problems [7], we solve some of them in this work, such as the identification of bigrams, trigrams or n-grams, which allows to compute the similarity between n-grams or sentences.

In our application, we only use the information provided by the organizers, which leaves us somewhat limited as to the visual concepts in the lifelog images. We believe that using the most recent state-of-art algorithms, a more rich description of the images can be obtained, resulting in a performance increase. In the future, we intend to integrate in our application features that have already been developed in previous work, such as selecting images in upload stage based on low level properties [8]. However, we think that using more of the metadata provided by the organizers can also improve the result. For example, make use of the GPS coordinates (latitude and longitude) to trace the lifelogger routes, such as the way home to work and vice-versa.

5 Conclusion and Future Work

The Lifelog Moment Retrieval (LMRT) sub-task of ImageCLEF lifelog 2020 was the baseline for a new web application that aims to help people to improve their quality of life.

We obtained the same exact results for the automatic approach (run 1 and run 2) even when using different state-of-the-art object detection algorithms and different weights for each category. Some of the reasons for this to occur is because much of the used information used was provided by the organizers, like activities and locations. Not only that, but the obtained scene recognition labels were not accurate enough.

In our interactive approach, using the application developed we were able to obtain a F1-measure@10 score of 0.52, which is till date our best. This makes us believe that an approach with visualization and user interaction is a more suitable method for a lifelog application. Although the results are already better compared to the previous work, our application is a baseline version which still requires improvements and new tools.

For future improvements in our approaches, we pretend to implement better scene recognition, object detection, activity and color detection algorithms, since color was a relevant element in some of the topics in the LMRT task. We will also use other data provided by the organizers, such as GPS coordinates and integrate features that have already been implemented in previous work.

6 Acknowledgments

Supported by the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010), co-funded by Centro 2020 program, Portugal 2020, European Union, through the European Regional Development Fund.

References

1. Dodge, M., Kitchin, R.: 'outlines of a world coming into existence': pervasive computing and the ethics of forgetting. Environment and planning B: planning and design **34**(3), 431–445 (2007)
2. Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017), to appear
3. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ştefan, L.D., Constantin, M.G.: Overview of the ImageCLEF 2020: Multimedia retrieval in lifelogging, medical, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
4. Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L.: Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 181–196 (2018)
5. Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., l Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T.: Overview of ImageCLEF Lifelog 2020:Lifelog Moment Retrieval and Sport Performance Lifelog. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)

6. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
7. Ribeiro, R., Neves, A.J., Oliveira, J.L.: Ua.pt bioinformatics at imageclef 2019: Lifelog moment retrieval based on image annotation and natural language processing. In: CLEF (Working Notes) (2019)
8. Ribeiro, R.F., Neves, A.J., Oliveira, J.L.: Image selection based on low level properties for lifelog moment retrieval. In: Twelfth International Conference on Machine Vision (ICMV 2019). vol. 11433, p. 1143303. International Society for Optics and Photonics (2020)
9. Wang, P., Sun, L., Smeaton, A.F., Gurrin, C., Yang, S.: Computer vision for lifelogging: Characterizing everyday activities based on visual semantics. In: Computer Vision for Assistive Healthcare, pp. 249–282. Elsevier (2018)
10. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)