DCU team at The 2019 Insight for Wellbeing Task: Multimodal personal health lifelog data analysis

Tu-Khiem Le[1*], Van-Tu Ninh[1*], Liting Zhou[1], Duc-Tien Dang-Nguyen[2], Cathal Gurrin[1]
[1] Dublin City University, Ireland
[2] University of Bergen, Norway
tukhiem.le4@mail.dcu.ie,tu.ninhvan@adaptcentre.ie,zhou.liting2@mail.dcu.ie,
ductien.dangnguyen@uib.no,cathal.gurrin@dcu.ie

ABSTRACT

In this paper, the authors described their proposed method in analyzing lifelog data in association with the environment. Tackling the problem of incomplete data, we proposed a replacement method using linear regression method which results in a normalized L2 distance score of 0.0153. Meanwhile, the authors solved the personal air quality subtask by inferring from lifeloggers' PM2.5 data, which achieves 1.0 in the arithmetic mean of absolute distance score between the predictions and the actual classes.

1 INTRODUCTION

Along with the development of engineering and technology, more and more personal devices such as smartphones, video cameras and wearable sensors have come to life which provide people the ability to easily capture every aspect of their life. On top of that, the term lifelogging is defined to be the process of recording a detailed trace of life passively[4], which generates a large collection of multimedia data. The huge amount of lifelog data leads to the need to quickly retrieve and extract particular insight based on the associations between data. In the MediaEval 2019 Insight for Wellbeing Challenge, they defined a new approach to lifelog data in relation with the environment. This is potential in analyzing the effect of general pollution on the living quality on individual scale. Beside the information recorded from the weather and air pollution stations, lifelog data could add in the true nature of particular regions where the stations are not set up.

The organizers generated a novel dataset called SEPHLA [6] which is collected by multiple lifeloggers who walk on several selected routes in the city and record data through wearable sensors and smartphones. The lifelog images, biometrics, weather, urban perception tags, emotional tags and air pollution data are provided within the dataset. To better understand the data and gain insights for personal wellbeing, the organizers defines two subtasks: Segment Replacement and Personal Air Quality prediction. In the first subtask, the participants are asked to investigate the associations among data and develop a solution to reconstruct the segments of data which are removed by the organizers. Meanwhile, The second subtask aims to estimate people wellbeing by predicting the AQI (Air Quality Index) on particular positions in a specified time. More details about the this challenge can be found in [5].

* These two authors contributed equally.

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

2 RELATED WORK

In recent years, lifelogging has gained more and more attentions and many research works have been proposed to provide better understanding of personal digital collections. To support, many international benchmarking efforts have been made and various challenges on lifelogging data were hosted, the most recent of which is NTCIR-14 Lifelog-3 Task [3], LSC 2018 [1], and ImageCLEF2019-lifelog [2]. While the purpose of these challenges is to mainly focus on developing a solution to retrieve relevant moments based on a set of given queries, each challenge has different subtasks to further explore this multimodal data. In the Lifelog Search Challenge (LSC), not only are the participants required to build an interactive retrieval system, but they also need to compete with each other in the competition with real-time on-screen query.

The datasets, which were utilised in these challenges, are collected by many lifeloggers who wear a passive-captured wearable camera and other tracking sensors. Each lifelogger normally generates around 1250 - 4500 images per day in association with other biometrics (e.g. heart rate, calorie), locations (GPS), physical movements and music. They share nearly the same structure with the lifelog data in the MediaEval 2019 Insight for Wellbeing Challenge. However, this challenge also considers additional information from the environment, which makes the insight more general and enables us to obtain an overview of the wellbeing among individuals.

3 APPROACH

From the dataset, we are provided air quality data gathered by the stations and lifeloggers' sensors. These are extremely useful information to reconstruct missing segments of data and predict air quality index for specific areas. Besides, we also got a collection of image data recorded by the lifeloggers with corresponding visual concepts extracted from the neural network, along with the information on the checkpoints where they are asked to take pictures. However, the images which are actively taken might vary from the lifeloggers' preferences. Therefore, it's hard to capture and generalize the context across individuals. Based on the observation we gained, we proposed the solutions to both sub-tasks which are described in the following subsections.

3.1 Segment Replacement

In this subtask, the sequence of missing PM2.5 data is specified in each query with a starting and ending time. As the lifeloggers walked in groups, the data from others could help regenerate the missing segments. The data from the stations, however, is not quite

reliable since their distance is too far from the routes and they might contribute noises to the result. Therefore, considering the NO2, O3, temperature, humidity and heartbeat data from people who share the same route with the targeted lifelogger, we build a simple linear regression model to predict the removed PM2.5 data. Specifically, let $\mathbf{x}$ be the 5-dimensional L2-normed feature vectors composed of five components mentioned above, and y be the targeted PM2.5 value that needs to be predicted. We construct a linear regression model $y = \mathbf{w}^T\mathbf{x} + b$ and apply gradient descent to find the best parameters $\mathbf{w}$ and b to minimize root-mean-square error, which aims to minimize the gap between model predictions and ground-truth of train data. Then the trained model is used to generate the missing PM2.5 of the targeted person in that group. As NO2, and O3 values are not almost zero for most of the times, temperature, humidity, and heart-beat are the main factors that contribute most to our predictions.

3.2 Personal Air Quality

To obtain AQI for each day, we would need to first gather the air quality data. From the checkpoints of each route, we could obtain a list of GPS along the route. Then, we extracted all air quality data where lifeloggers' GPS is closed to the checkpoints. The distance between two GPSs is calculated using the Haversine formula.

As we observed from the air quality data of each route, NO2 and O3 values are mostly zeros while PM2.5 values have some fluctuations. Therefore, we choose PM2.5 to predict the ultimate Air Quality Index (AQI). At first, we refine the data to get the right PM2.5 data for each route by calculate the distance between the route's GPS and collectors' current GPS. After this step, we obtain data for 27 routes on 7 days from different groups of collectors. For each data on a day collected by a user, we compute its average PM2.5. Therefore, we receive many average PM2.5 values from many collectors in one day. We consider the maximum value of these average PM2.5 values as the criteria to evaluate AQI for that route on that day. Then, we average the AQI value of 7 days and re-evaluate again to infer the AQI level of the route.

4 RESULTS AND ANALYSIS

Table 1: Ranked list of best score of each team in Segments Replacement Subtask

Group ID	Run ID	Score
healthism	3	**0.000427182**
SHT-UIT	3	0.000463205
DCU	**1**	**0.015310414**
HCMUS	4	0.015514208

It can be seen from the table 1 that our team (DCU) manages to achieve the 3^{rd} highest score of approximately 0.0153 among the best submission list in the Segment Replacement sub-task. It means that our approach manages to generate relatively good prediction with low error. However, there are other solutions could provide more precise result with significantly low error.

Meanwhile, in the Personal Air Quality sub-task, our approach got the arithmetic mean absolute L1 distance score of 1.0. This means that our approach to handle the data for this task is not good and the operation that we apply to process PM2.5 data to infer AQI level is not correct. Since the data recorded from the lifeloggers walking through the route is not totally correct (as the values are almost zeros for all) and the collected data is not enough (less than 24 hours during seven non-consecutive days), we can hardly infer the right AQI level for the route.

Table 2: Ranked list of best score of each team in Personal Air Quality Subtask

Group ID	Run ID	Score
healthism	19	**0.3**
SHT-UIT	1	0.8
DCU	**1**	**1.0**

As we do not exploit all the provided materials such as the data recorded from the stations, images and related metadata, we might miss some important features that could be used to improve our predictions. Moreover, as we rely on the users' recorded data along the route that they pass through, the recorded values such as PM2.5, NO2, O3 are not reliable as the most of their values are zeros. These are the main factors that affect our results in both sub-tasks. In order to improve it in future work, we might need to consider additional data on the internet, which is recorded from nearby stations, to provide the missing PM2.5 values during the days to generate the correct estimation of AQI score.

ACKNOWLEDGMENTS

This publication has emanated from research supported in party by research grants from Irish Research Council (IRC) under Grant Number GOIPG/2016/741 and Science Foundation Ireland under grant numbers SFI/12/RC/2289 and 13/RC/2106.

REFERENCES

[1] 2018. *LSC '18: Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge*. ACM, New York, NY, USA.

[2] 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.

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

[4] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Foundations and Trends® in Information Retrieval* 8, 1 (2014), 1–125. https://doi.org/10.1561/1500000033

[5] Tomohiro Sato Koji Zettsu Duc-Tien Dang-Nguyen Cathal Gurrin Ngoc-Thanh Nguyen Minh-Son Dao, Peijiang Zhao. 2019. Overview of MediaEval 2019: Insights for Wellbeing Task: Multimodal Personal Health Lifelog Data Analysis. In *MediaEval2019 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Sophia Antipolis, France.

[6] Tomohiro Sato, Minh Dao, Kota Kuribayashi, and Koji Zettsu. 2018. SEPHLA: Challenges and Opportunities within Environment-Personal Health Archives.