Building a Disclosed Lifelog Dataset: Challenges, Principles and Processes

Duc-Tien Dang-Nguyen
Insight Centre for Data Analytics
Dublin City University
duc-tien.dang-nguyen@dcu.ie

Liting Zhou
Insight Centre for Data Analytics
Dublin City University
zhou.liting2@mail.dcu.ie

Rashmi Gupta
Insight Centre for Data Analytics
Dublin City University
rashmi.gupta3@dcu.ie

Michael Riegler
Simula Research Laboratory
University of Oslo
michael@simula.no

Cathal Gurrin
Insight Centre for Data Analytics
Dublin City University
cgurrin@computing.dcu.ie

ABSTRACT

In this paper, we address the challenge of how to build a disclosed lifelog dataset by proposing the principles for building and sharing such types of data. Based on the proposed principles, we describe processes for how we built the benchmarking lifelog dataset for NTCIR-13 - Lifelog 2 tasks. Further, a list of potential applications and a framework for anonymisation are proposed and discussed.

CCS CONCEPTS

•**Information systems** → **Information retrieval; Specialized information retrieval;** •**Security and privacy** → *Privacy-preserving protocols;*

KEYWORDS

Lifelogging, visual lifelog, dataset

ACM Reference format:
Duc-Tien Dang-Nguyen, Liting Zhou, Rashmi Gupta, Michael Riegler, and Cathal Gurrin. 2017. Building a Disclosed Lifelog Dataset: Challenges, Principles and Processes. In *Proceedings of CBMI '17, Florence, Italy, June 19-21, 2017,* 6 pages.
DOI: 10.1145/3095713.3095736

1 INTRODUCTION

Within the last decade, we have witnessed the "explosion" of wearable devices, with 102.4 million devices shipped in 2016[1]. Using these small devices, we can passively capture pictures, videos, audio, and also biometric information in every moment of our life, and thus, we are creating vast archives of personal data where the totality of an individual's experiences are stored permanently as a personal multimedia archive. These unified digital records, commonly referred to as *lifelogs* [1], are unprecedented in terms of volume and variety. Shown in Figure 1 is the data of a logger gathered during 10 years, which significantly increases from less than 5 GB with only images and activities logs in 2005 to more than 3.2 TB with rich multi-modal information from images, audio, videos to biometrics in 2015. Captured over a long period of time (e.g., a decade for the logger whose data is reported in Figure 1), heterogeneous lifelogs gathered increasing attention in recent years within the research community to provide a detailed picture of the experiences of an individual, with numerous applications in terms of assisted technologies for human memory [2], health and wellness [3, 4], activities recognition [5], and many others. It is no surprise that lifelogging is also receiving increasing attention within the research community and is fast becoming a mainstream research topic with the increase of workshops focusing on lifelogs, e.g., NTCIR-12 - Lifelog[2], ACM MM 2016 - LTA[3], ImageCLEF 2017 - Lifelog[4] and NTCIR-13 - Lifelog 2[5].

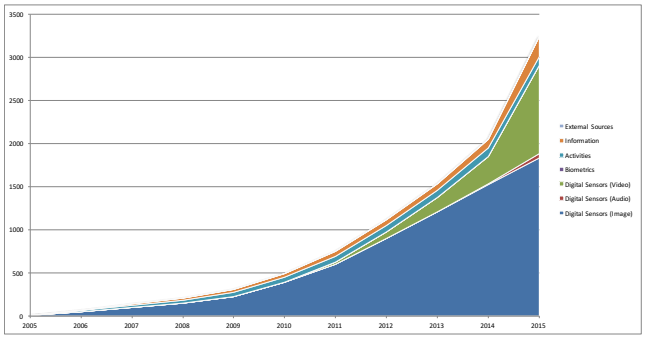


Figure 1: The amount of the lifelog data of a logger significantly increases over 10 years.

In order to get insights from the lifelogs, individuals or researchers require tools for knowledge extraction, search, summarisation, and visualisation. Moreover, to validate the performance of these tools, they need data, including large and rich collections of lifelog data. Therefore there is a strong need to have a common published dataset for studies in this field [1]. However, the design and construction of a lifelog dataset that will be shared and studied in public is not

[1] According to CNET, http://bit.ly/wearable2016

[2] http://research.nii.ac.jp/ntcir/ntcir-12/index.html.
[3] http://lta2016.computing.dcu.ie/
[4] http://imageclef.org/2017/lifelog
[5] http://ntcir-lifelog.computing.dcu.ie/

CBMI '17, Florence, Italy

DOI: 10.1145/3095713.3095736

trivial, where there are significant technical challenges to be solved, arising from the gathering, semantic enrichment, and pervasive accessing of these vast personal data archives [6]. To the best of our knowledge, there has been only a published lifelog dataset that fulfilled major of the problems above: the NTCIR-12 - Lifelog [6, 7]. The challenges of building a disclosed lifelog dataset can be summarized as follows:

- **A willingness to share?** The success of the quantified-self movement [8] shows the willingness of individuals to gather archives of their life. People tend to familiar with gathering their personal data, or even being a subject captured in a lifelog of someone else. However, finding people who are willing to donate years or even decades of data is a challenge.
- **What to log?** Following the spirits of the definition by Dodge and Kitchen [9], a lifelog typically consists of numerous different types of data, such as image/video content from wearable cameras (e.g., SenseCam), audio content from personal audio devices, biometric sensor content from activity trackers (e.g., from a wristband or by a phone as in [5]) or health-monitoring devices, informational content from the media consumed by the lifelogger, and so on. Ideally, we should log all information from all sources, however, it is not doable in practice. Thus, making a decision on what to log is indeed a non trivial task.
- **How often to log?** Lifelog data shows considerable variance in terms of capture velocity and type variety. Some sensors, such as biometric sensors can capture data on a second-by-second basis, whereas wearable cameras may capture between 1 to 5 images per minute. In order to be useful for the individual, lifelogging needs content organisation and retrieval facilities that operate over data at different velocities and frequencies, in order to address a wide variety of use-cases.
- **What NOT to be shared?** Privacy and data security, which has implications for both the individual and society as a whole [10], are also important issues that need to be considered. Personally identifiable information, e.g., personal ID numbers, car plates, addresses, and others, should not be shared. This raises the challenge of how to filter out such information in order to protect the lifeloggers as well as individuals that appear in the lifelogs.
- **Who can access the data?** The goal of a published dataset is to share with researchers. Lifelogs, however, differ from more traditional shared datasets [6], e.g., the privacy-aware issues, and thus need a well designed strategy to control data access and modify (if needed) the lifelog data.

In this study, we aim at proposing principles for construction of a disclosed lifelog dataset as well as describe the process of how we built a new lifelog dataset, the NTCIR-13 - Lifelog 2[6], following the proposed principles.

The contributions of this paper are: (i) we point out the challenges for building a shared lifelog dataset; (ii) we propose principles for building a shared lifelog dataset (Section 2); (iii) based on the proposed principles, we build and describe the whole processes from data gathering to determine the roles for the people who are building, sharing and exploiting a disclosed lifelog dataset (Section 3); (iv) a framework for anonymisation (Section 4); and (v) a short list of potential applications are recommended (Section 5).

[6]Lifelog 2 is the core task of NTCIR-13 aims to advance the state-of-the-art research in lifelogging.

2 PRINCIPLES OF BUILDING A LIFELOG DATASET

Learning from the NTCIR-12 [6, 7] and the LTA-2016, we propose these principles for building a disclosed lifelog dataset:

1. The Continuity. *Lifelog data of each individual should be captured continuously for at least 15 days.*

Getting insights from lifelogs could be a retrieval problem, e.g., *"find the moments that a logger having dinner"*, or could be also an analytics problem, e.g., *"compare the eating habits between two loggers"*. In order to answer these queries, the dataset must contains sufficient information, i.e., the lifelogs of each logger should be gathered continuously over a number of days. In this study, we propose that the minimum amount should be 15 days, which we learned from NTCIR-12, that a 15-day period is sufficient for identifying habit patterns of an individual.

2. The Completeness. *The lifelogs should contain four basic types of information: visual data, personal biometrics, human activity, and information accesses.*

In order to allow for statistically significant studies in the field of lifelogging, a dataset needs to be large enough to represent real-world data of lifeloggers [6]. As addressed in the challenges, we should log all information from all sources, however, it is not doable in practice, thus, we propose these four sources as that main data that a lifelog dataset should contain: **(i) Visual information:** use for the continuous and automatic capture of life activities as a visual sequence of digital images. **(ii) Personal biometrics:** the lifelogs should contain basic personal biometrics of the individual, e.g., heart-rate, calorie, and others. **(iii) Human activity:** There are many personal sensing devices for monitoring everyday activities, which can monitor human performance, for example activity levels (number of steps, distance traveled), sleep duration, etc. **(iv) Information access:** This could refer to using a computing device (e.g. smartphone or computer) to continuously and passively capture the userfis context or logging the consumed data from all the words typed, web pages read, YouTube videos watched and so on. It worth noticing that the NTCIR-12 dataset does not have personal biometrics and information access, and thus it does not fully satisfy this principle.

3. The Anonymity. *All user-identifiable data have to be removed.*

We must consider the principles of privacy-by-design when creating the lifelog dataset. In order to remove user-identifiable information while maintain the usefulness of the data, we propose these information have to be anonymised: human faces, personal IDs, vehicle plates, name tags, and the address/location of the individual's home and work place.

4. The Protectiveness. *The dataset should be password protected and all accesses should be logged. The lifeloggers who donated the data, or any identifiable individuals who appear in the data can request some content to be deleted at any time and need to agree to use of data beyond any initial planned dataset release. The researchers who manage the dataset can also request data deletion at any time.*

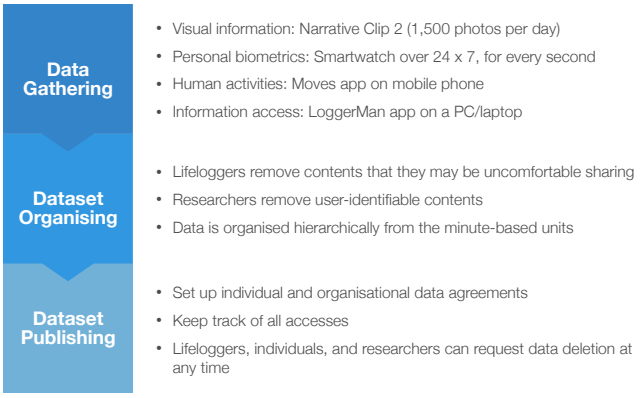


Figure 2: The proposed process for building a disclosed lifelog dataset.

The lifelog data is anonymised, however, it contains experiences of loggers as well as individuals appeared in the lifelogs, and thus it must be only shared for research or education purposes. These information, hence, should be also protected strictly by agreements between loggers, users and the people who manage the dataset.

These principles answer the last four challenges pointed in Section 1 while we ourselves and some other volunteers provide the solution for the first challenge. Applying these principles, we collected the lifelog data which gathered continuously for over 45 days from rich multi-modal sensors and devices, and built a new disclosed lifelog dataset: the NTCIR-13 - Lifelog 2. In the next section, we describe the process for building this dataset. It is worth noticing that following these principles, NTCIR-13 - Lifelog 2 is reusable, which means it can support a number of years of ongoing research activities.

3 PROCESSES

In order to build a lifelog dataset that respects the proposed principles, we apply a 3-step process, summarized in Figure 2, as follows: firstly we gather individuals' lifelogs by following the first two principles. Next, the gathered data is cleaned by removing user-identifiable data (applying the anonymity principle) and then organized hierarchically with the basic units composed from every minute. Finally, before publishing the dataset, we protect the data by making agreements between loggers, users and the people who manage the dataset, as well as putting in place a data download tracking mechanism.

These steps are described as follows:

Step 1. Data Gathering

Due to the long-term, always-on, nature of lifelog data gathering, it was important to reduce the overhead on the lifelogger of gathering the data. Hence, the data was gathered using only their mobile phones with two wearable devices: a Narrative Clip 2[7] and a smartwatch[8] (see Figure 3). It is worth noticing that in the NTCIR-13 - Lifelog 2 dataset, each individual gathered continuously over 45 days, that fulfilled our first principle.

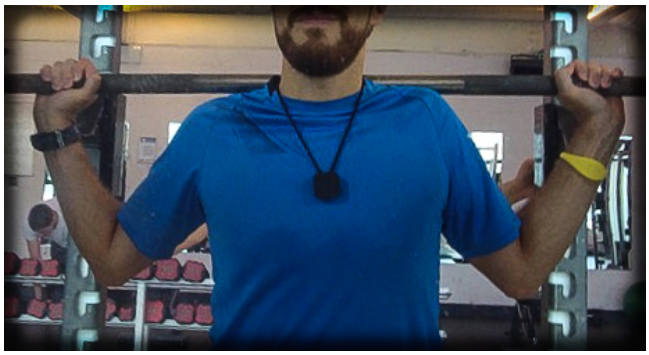


Figure 3: A lifelogger with his everyday wearable devices: A Narrative 2 camera (wearing on his neck) and a smartwatch (the black watch on his right arm).

The data was gathered as follows:

- Visual information: the visual data was gathered using a Narrative Clip 2, taking photo (landscape whenever possible) at each 45 second, from breakfast to sleep. This produces about 1,500 images per day.
- Personal biometrics: We use the smartwatch to obtain the biometrics information over 24×7, gathering these information every second: heart rate, galvanic skin response, calorie burn, and steps. Beside such biometric information, some lifeloggers also provide their blood pressure and blood sugar, measured every morning.
- Human activity: In order to log the individuals physical activities, the Moves app[9] is used. Moves app is a smartphone app that automatically records user activity in terms of semantic locations and physical activities (e.g., waking, running, transport), without requiring any user intervention. This app was installed on the personal smartphones used by the lifeloggers. The moves data was exported from the Moves cloud-service after the data gathering process was complete.
- Information access: Is collected using the LoggerMan [11] app. LoggerMan helps researchers and lifeloggers to collect interaction data produced during normal computer usage. The main goal of LoggerMan is to work passively in the background, intercept usage events and store them for later analysis. It gathers wide range of keyboard, mouse and UI actions, thereby capturing the information creation and consumption activities of the individual.

Step 2. Dataset Organising

From the gathered the data, the second step is applied to first remove user-identifiable content, i.e., apply the anonymisation process (details of this process is described in the next section).

The anonymised data is then checked to ensure temporal alignment of the sensor data, since it is gathered from different devices. It was necessary to check and resolve alignment problems (typically in the order of 1-2 minutes) for a lifelogger by cross-referencing

[7] http://getnarrative.com/
[8] The lifeloggers used the BASIS smartwatch.
[9] https://moves-app.com/

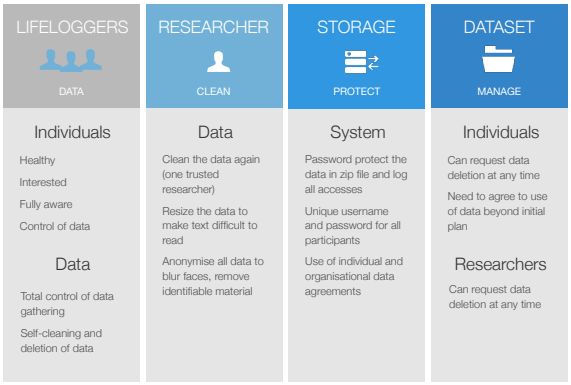


Figure 4: The components of a published lifelog dataset.

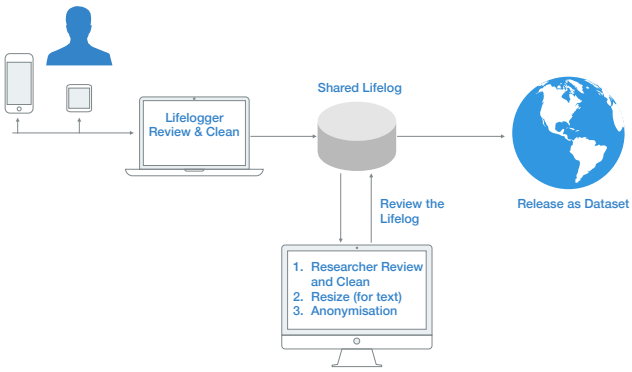


Figure 5: The anonymisation process applied on raw data.

reported timestamps from the Narrative 2 camera with clocks captured daily in the real-world.

Finally, the synchronized data is organised hierarchically from the minute-based units. Typically in information retrieval (IR), there is a single basic unit for indexing and retrieval. For many IR tasks, this basic unit is the preferred as unit of retrieval and choosing the basic unit may not trivial. In web search, a document could be considered as a basic unit, however, for other IR problems, a basic unit as a document is too large to be of value to the user. Consider video search as an example of an IR application, a video can be seen as too large to be useful, and thus, the video is often segmented into sub-units called video shots or scenes which are shorter and more likely to answers a user's information need. Back to the lifelog data, it is not trivial to decide what the basic unit is since the lifelogs are composed of different types of data, captured at different frequencies (1 second to potentially 1 day). In order to deal with this problem, we propose to organize the data hierarchically where the top levels are the days, sorted by chronological order, and each day contains $1,440$ (24 hours $\times$ 60 minutes) minute-based units. Verified by the NTCIR-12 task, it is confirmed that this structure is useful for analytics approaches to get insights from the lifelog data.

Step 3. Dataset Publishing

For dataset publishing, we set up agreements between lifeloggers, users, and researchers/people who manage the dataset and keep track for all accesses. We also define four components (summarized in Figure 4) of the published dataset as follows:

(1) Lifeloggers: provide the data and also do the self-cleaning on the data. They (the lifeloggers) also take the responsibility of letting the individuals in their field of view be fully aware that they (the individuals) are being captured by the camera.
(2) Researcher/People who manage the dataset: do the anonymisation and organise the data.
(3) Storage: Beside hosting the dataset, this "component" also protects the data by applying password protection for each data file. We also protect the data by making individual and organisational data agreements.
(4) Dataset: the lifeloggers and individuals can request data deletion at any time and need to agree to use of data beyond initial plan. The researchers can also request data deletion at any time.

4 ANONYMISATION

How to support sharing and privacy-aware analytics is a major issue in lifelogging, and thus, we use this whole section to discuss this issue as well as describe the anonymisation process applied in building NTCIR-13 - Lifelog 2 dataset.

4.1 Data Level

Lifeloggers who donate their dataset are willing to share their everyday activities, however, it was necessary to give them an opportunity to remove any data that they may be uncomfortable sharing. This elaborated a manual inspection of all their lifelog data before sharing it with the researchers. After this, all images were reviewed by a trusted researcher with oversight of the entire dataset to ensure that no potentially embarrassing or offensive images were accidentally included in the dataset.

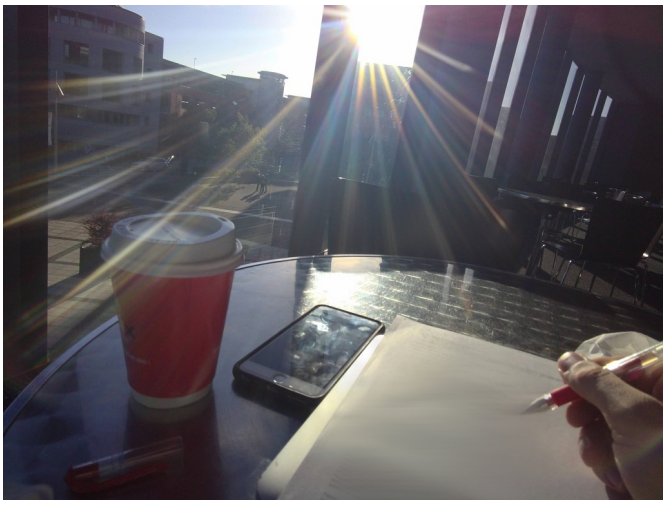
To ensure privacy of both the lifeloggers and individuals (subjects and bystanders) captured in the lifelog, user-identifiable content is removed by the following process (summarized in Figure 5 where some visual anonymisation results can be seen in Figure 6): First, each recognisable face in every image was blurred in a manual process, which ensured no false positives or missed faces. It is worth noticing that this process has to be done manually since the current face detection methods are not capable of detecting every human face in the wild. Car plates, handwritten notes, or any other user-identifiable contents are also manually blurred from the images. Next, the images are resized down to 1024×768 resolution, in order to reduce the storage space as well as making the majority of any on-screen text captured by the lifelogging camera unreadable.

Similar rules are applied on the information access and creation data collected by LoggerMan, all user-identifiable contents are removed from the information access data.

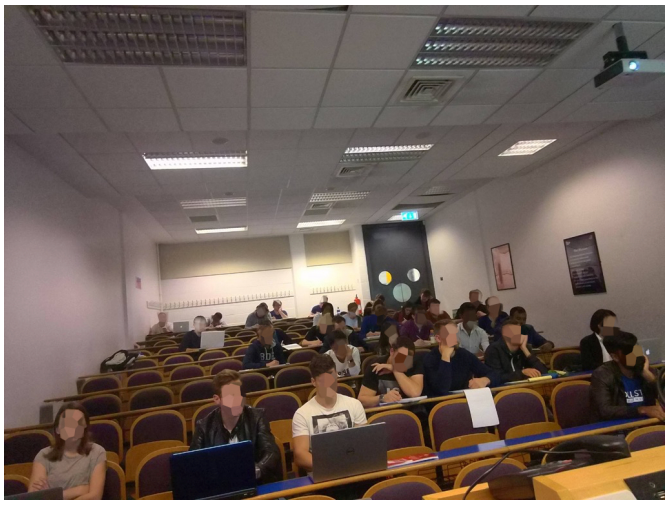
Since privacy extends beyond content of images, the Moves app automatically converts all locations from absolute GPS locations into semantic locations, which resulted in potentially sensitive absolute addresses being labeled with generic names such as *'homefi* or *'work'*, thereby making it more unlikely that the lifeloggers could be identified.

(a) A close-up human face and his name tag.

(b) A car plate.

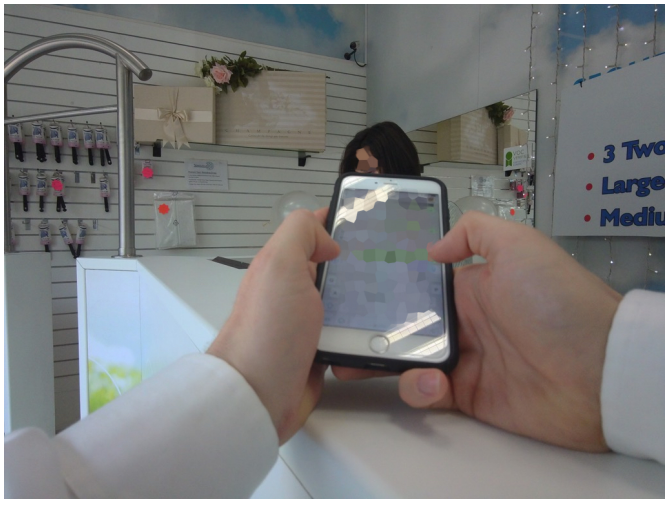


(c) Handwritten notes.



(d) Human faces.

(e) Monitor.



(f) Contents on mobile screen.

Figure 6: Examples of the visual anonymisation.

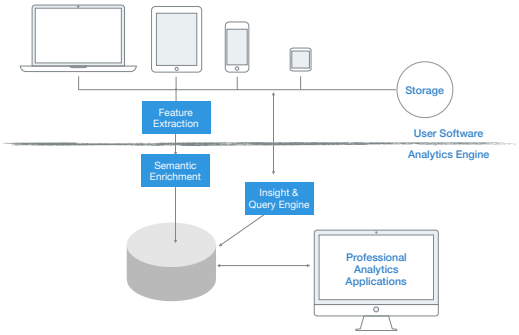


Figure 7: A proposed privacy-aware lifelogging model.

4.2 Application Level

Beside the anonymisation at data level, we also propose a model for protecting the privacy at the application level, i.e., allow doing analytics without working on the raw data. This is done by a classic way (summarized in Figure 7): providing meaningful features instead of the raw data. Among different types of features, we decided to exploit the Microsoft Computer Vision API[10] to extract the information about visual content found in the lifelog images. This API allows to identify content and label it with confidence, which allows different approaches can get insights from the lifelogs without working on the raw data. These semantic features are published together with the images in the NTCIR-13 - Lifelog 2 dataset.

5 POTENTIAL APPLICATIONS

Having a large collection of annotated personal data for a long time, such as this dataset, opens up a number of research opportunities. In this section, we discuss some of potential applications which can be built by exploiting undisclosed lifelog datasets.

This dataset can be served as a benchmarking initiative for lifelogging studies. A typical issue in research communities is the lacking of common dataset and evaluation methods. Thus, the nature purpose of these lifelogs, is serving the research community as a common benchmarking initiative, where queries can be built based on the lifelogs, together with the ground-truth annotations, e.g., building as a retrieval task as in NTCIR-12 - Lifelog task [6, 7]. Then, studies in lifelogging can use these queries and annotations for evaluations or comparisons with other methods.

Generating insights & analytics from real-world datasets of wearable personal sensing. This is approved by the insights task in NTCIR-12, where participants were asked to provided insights from the lifelogs data.

With a rich information from different sensors, a typical application is multi-modal search and retrieval over archives of personal data.

Lifelog-specific applications: addressing many applications of memory and personal health care. There are some lifelog application examples in the research of health care, e.g., the UbiqLog [4] application which is a custom-build every new lifelogging application using on smartphones.

[10]https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

Visual concept extraction, from real-world all-day wearable camera data (with additional supporting meta-data). By extracting visual concepts, we can see many potential applications by exploiting the lifelogs in the dataset. For example, recognising what a lifelogger eats everyday and his/her biometrics information, then recommendations for his eating habits could be given, e.g., should he/she change their ways of eating to avoid diabetes or obesity. It is gathering and managing health lifelog data for improving of the quality of person's life.

Activity recognition utilizing wearable sensors has attracted researchers for years. These approaches aim to analyze data gathered from wearable devices to semantically describe human activities. Thus, this lifelog dataset will contribute valuable data for researchers in this field.

With the principles of the continuity and completeness, this lifelog dataset can be considered as ideal resources for time-series analysis research. For example, from all-day personal data over extended time-periods from one, or multiple users, the human activities can be recognized by exploiting time-series analysis.

Privacy-aware retrieval, to explore the privacy concerns surrounding search and retrieval from large lifelog archives.

6 CONCLUSION

In this paper we reported challenges and proposed principles for building a disclosed lifelog dataset. This is the first time the problems of building such kind of dataset has been deeply discussed. As a demonstration, a full process for building a published lifelog dataset - the NTCIR-13 - Lifelog 2 was described in details. A list of potential applications and a framework for anonymisation were also proposed and discussed. In future work, we will extend the data with more information and annotations, making it become a valuable resources for lifelogging research communities.

REFERENCES

[1] C. Gurrin, A. F. Smeaton, and A. R. Doherty, "Lifelogging: Personal big data," *Found. Trends Inf. Retr.*, vol. 8, no. 1, pp. 1–125, Jun. 2014.
[2] S. Hodges, E. Berry, and K. Wood, "Sensecam: A wearable camera that stimulates and rehabilitates autobiographical memory," *Memory*, vol. 19, no. 7, pp. 685–696, 2011, pMID: 21995708.
[3] M. B. Amin, O. Baños, W. A. Khan, H. S. M. Bilal, J. Gong, D. Bui, S. H. Cho, S. Hussain, T. Ali, U. Akhtar, T. Chung, and S. Lee, "On Curating Multimodal Sensory Data for Health and Wellness Platforms," *Sensors*, vol. 16, no. 7, p. 980, 2016.
[4] R. Rawassizadeh, M. Tomitsch, K. Wac, and A. M. Tjoa, "Ubiqlog: a generic mobile phone-based life-log framework," *Personal and ubiquitous computing*, vol. 17, no. 4, pp. 621–637, 2013.
[5] M.-S. Dao, D.-T. Dang-Nguyen, M. Riegler, and C. Gurrin, "Smart Lifelogging: Recognizing Human Activities using PHASOR," in *International Conference on Pattern Recognition Applications and Methods*, 2017, pp. 761–767.
[6] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal, "NTCIR Lifelog: The First Test Collection for Lifelog Research," pp. 705–708, 2016.
[7] ——, "Overview of NTCIR-12 Lifelog Task," pp. 354–360, 2016.
[8] M. Swan, "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery," *Big Data*, vol. 1, no. 2, pp. 85–89, Jun. 2013.
[9] M. Dodge and R. Kitchin, "'Outlines of a World Coming into Existencefi: Pervasive Computing and the Ethics of Forgetting," *Environment and Planning B: Planning and Design*, vol. 34, no. 3, pp. 431–445, 2007.
[10] K. O'Hara, M. M. Tuffield, and N. Shadbolt, "Lifelogging: Privacy and empowerment with memories for life," *Identity in the Information Society*, vol. 1, no. 1, pp. 155–172, 2008.
[11] Z. Hinbarji, R. Albatal, N. E. O'Connor, and C. Gurrin, "Loggerman, a comprehensive logging and visualisation tool to capture computer usage," in *22st International Conference on MultiMedia Modelling*, 2016, pp. 342–347.