CHAPTER 1

Computer Vision for Lifelogging [1]

Characterizing Everyday Activities Based on Visual Semantics

Peng Wang[2,a,*], Lifeng Sun[*], Alan F. Smeaton[**], Cathal Gurrin[**] and Shiqiang Yang[*]

* National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Haidian District, China

** Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland

[a] Corresponding: pengwangnudt@sina.com

Abstract

The rapid development of mobile devices capable of sensing our interaction with the environment has make it possible to assist humans in daily living such as helping patients with cognitive impairment or providing customized food intake plan for patients with obesity, etc. All of this can be achieved through the passive gathering of detailed records of everyday behaviour which is termed as lifelogging. For example, the widely adopted smart mobiles and newly-emerging consumer wearable devices like Google glass, Baidu eye, Narrative clip, etc. are usually embedded with rich sensing capabilities including camera, accelerometer, GPS, digital compass, etc. which can help to capture daily activity unobtrusively. Among such heterogeneous sensor readings, visual media contain more semantics to assist in characterizing everyday activities and visual lifelogging is a class of personal sensing which employs wearable cameras to capture image or video sequences of everyday activities. This chapter will focus on the most recent research methods in understanding visual lifelogs, including semantic annotations of visual concepts, utilization of contextual semantics, recognition of activities, visualization of activities, etc. We also discuss some research challenges which indicates potential directions for future research. This chapter is intended to support readers in the area of assistive living using wearable sensing, computer vision for lifelogging, human behaviour researchers aiming at behavioural analysis based on visual understanding.

Chapter points

- Comprehensive description of lifelogging and visual lifelogging in assistive living.
- State-of-the-art computer vision processing for visual understanding everyday contexts.
- A fully understanding of everyday activities from static to dynamic point of views.

[1]This chapter is supported by Natural Science Foundation of China under Grant No. 61502264, 61210008, Beijing Key Laboratory of Networked Multimedia, and Science Foundation Ireland under grant number SFI/12/RC/2289.

[2]This is author footnote

- Practical experience and guidance in visual lifelogging interpretation.

1. Introduction and Background

The proliferation of modern sensing devices have opened the possibility for technically assisted living which benefits the well-being of dependents such as the elderly, patients in need of special care, and independently healthy people. In order to do so, it is imperative to understand everyday activities of the individual in order to provide customized services or treatments. For example, the accurate sensing of the Activities of Daily Living (ADL) has many benefits, such as in analyzing human lifestyle, diet monitoring, occupational therapy, aiding human memory, active rehabilitation, etc. This derives the phenomenon of *lifelogging*, which is used to describe the process of automatically, and ambiently, digitally recording our own day-to-day activities for our own personal purposes, using a variety of sensor types.

1.1. Lifelogging in General

Lifelogging is a very broad topic both in terms of the technologies that can be used, as well the applications for lifelogged data. Compared to traditional digital monitoring through which users are monitored by others, lifelogging introduces a new form of sousveillance, i.e., capturing data about oneself for use by oneself [1, 2]. This is opposed to having somebody else record what we are doing and using the logged data for some public or shared purpose [3]. Among various applications of lifelogging, assistive living accounts for an important part of lifelogging research aiming at improving the health and well-being of human both mentally (such as memory recall) and physically (such as anomaly detection), thanks to the advantages of lifelogging in measuring activities longitudinally in fine granularity.

1.1.1. Context Sensing for Lifelogging

As an integrated part of our lives, our context is changing dynamically and if we can capture some parts of this context then these can be used as cues for reflecting our activities. By "context" we mean the features of where we are, who we are with, what we are doing and when we are doing it. Since the context includes various aspects of the environment in which the user interacts with digital devices, the plurality of context can be applied intelligently to detect meaningful changes in the environment for assistive living for example. The increasing adoption of sensors makes it possible to gather more context information on mobiles or wearable devices which is an important data source for activity recognition and understanding . This kind of applications of heterogeneous sensors in context sensing is named as multimodel context-awareness.

Based on the collection of low-level sensor information we can infer cues about the host and the environment.

The earliest motivation behind automatic context recording and generation of personal digital archives can be traced back to 1945 when Bush expressed his vision [4] that our lives can be recorded with the help of the technology and the access can be made easier to these 'digital memories'. This idea is unprecedentedly acceptable in the era of mobile sensor network when the cost of integrated sensors, storage and computational power is much lower. However, there is still a lack of consensus for a definition of lifelogging. In this chapter, we borrow the definition from [5] as:

BOX 1.1 Definition of Lifelogging

Lifelogging is the process of passively gathering, processing, and reflecting on life experience data collected by a variety of sensors, and is carried out by an individual, the lifelogger. The corresponding data gathered in lifelogging is termed as *lifelog*, which could be in heterogenous formats such as videos, pictures, sensor streams (like GPS locations or accelerometer traces).

From this definition, we can find that the accurate quantification of human activities using lifelogging can help to measure our diets, entertainments, leisures and sports, etc., more effectively. The longitudinal profiles in terms of digital media can provide better ways to record, analyze, understanding and further improve ourselves. Such digitally recorded contexts usually compensates the subjectivity of human feelings which tends to be limited by human intuition. For example, subjects with obesity often underreport their intakes and this significantly limits the food recording method [6, 7].

Context metadata like date, time and location may be sufficient for many lifelogging applications but there are others which require searching through lifelogs based on visual content, and for this to happen the automatic recording of visual inputs needs to be introduced. When visual sensing devices such as digital cameras or camera-enabled mobile devices are involved in the recording, we refer to such lifelogging as visual lifelogging, as defined in Box 1.2. Because visual information contains more semantics of events which can be used to infer other contextual information like 'who, what, where and when', visual lifelogging can usually act as the prosthesis of the lifeloger's experience and this forms a stream of lifelogging research and assistive living.

BOX 1.2 Visual Lifelogging

Visual lifelogging represents a branch of lifelogging research in which digital cameras or

camera-enabled mobile devices are employed as sensors to gather visual media to reflect the interaction of the lifelogger and his/her environment.

As the enabler for visual lifelogging, camera-enabled sensors are used in wearable devices to record still images [40] or video [3,13,32] taken from a first-person view, i.e. representing the subject's view of everyday activities. Visual lifelogging has already been widely applied in assistive living applications including aiding human memory recall, diet monitoring, chronic disease diagnosis, recording activities of daily living and so on. Example visual lifelogging projects include Steve Manns WearCam [31,32] , the DietSense project at UCLA [38] , the WayMarkr project at New York University [4] , the InSense system at MIT [3] , and the IMMED system [33] . Microsoft Research catalysed research in this area with the development of the SenseCam [12,40] which was made available to other research groups in the late 2000s.

1.1.2. Visual Lifelogging Categories

In terms of sensing devices, assistive viusal lifelogging can be categorized roughly into in-situ and wearable lifelogging.

In-Situ Visual Lifelogging

In-situ lifelogging can be described simply as sensing in instrumented environments such as homes or workplaces. This means that human activities can be captured through sensors such as video cameras installed in the local infrastructure, therefore the recording is highly dependent on instrumented environments, such as PlaceLab (MIT) [8]. Typical use of video sensors for in-situ sensing also includes works as reported in [9, 10, 11, 12] and [13] . Jalal et al. [11] proposed a depth video-based activity recognition system for smart spaces based on feature transformation and HMM recognition. Similar technologies are applied in other work by the same authors in [9] which can recognize human activities from body depth silhouettes. In related work by Song et al. [12], depth data is utilized to represent the external surface of the human body. By proposing the body surface context features, human action recognition is robust to translations and rotations. As with Jalal's work in [10], Song's work [12] still depends on static scenes with an embedded sensing infrastructure. Current activity recognitions in such lifelogging settings usually assume there is only one actor in the scene and how these solutions can scale up to more realistic and challenging settings such as outdoors are difficult.

Wearable Visual Lifelogging

Because frequent in-situ observations for activity measuring are are limited in the instrumented environments, the scaling up to more realistic and challenging settings

such as outdoors are difficult. The wildly adoption of mobile or wearable devices makes it feasible to measure everyday activities digitally with their built-in sensing and computational capabilities, which helps to alleviate the challenges of in-situ sensing. Meanwhile, activity recognition within such non-instrumented environments using wearable visual sensing is also a focus of assitive sensing and independent living. In wearable assistive living, the sensing devices are portable and worn directly by the subjects and can include head-mounted cameras in works by Hori and Aizawa [14] and Mann et al. [15] or cameras mounted on the front of chests in works by Blum et al. [16] and by Sellen et al. [17]. These literatures all reflect a common phenomena of continuous recording of everyday activity details based on wearable sensing. In this chapter, without explicit specification, we refer wearable lifelogging to lifelogging for simplicity purpose.

1.2. Typical Applications in Assistive Living

As we introduced previously, the development of sensors and low-cost data storage have make the continuous or long-term lifelogging possible. However, this is only the prerequisite of lifelogging because the necessary condition of lifelogging includes various applications which motivate the lifeloggers to do so. Gordon Bell, a scientist in Microsoft, is a senior lifelogger and attempted to digitalize the lifetime archives including articles, letters, photos, medical recordsin MyLifeBits project [18]. In his coauthored book Total Recall [19], he envisioned the roles of lifelogging in changing human daily life from various aspects like study, work, domesticity, etc. In this book, he also mentioned Cathal Gurrin, one author of this chapter, started to wear a SenseCam from mid-2006 to gather an extensive visual archive of everyday activities. According to Gurrin's experience, the passively captured visual lifelogs constructed a surrogate memory to re-experience his episodes of interest [19], such as when and where he met with an important person for the first time.

These overseen or experienced superiority of visual lifelogging all benefits from its detailed sensing, high storage, and multimedia presentation capabilities. Due to these advantages, visual lifelogging can be embraced in assistive living to satisfy the needs of different groups. The typical applications can be summarized as memory aid, diet monitoring, for ADL analysis, or disease diagnosis, and so on, though new application areas are emerging.

Memory aid: Memory aid is a potential medical benefit which can be supported by lifelogging technologies. By recording various aspects about our recent daily activities, visual lifelogging will offer an approach for wearers to re-experience, recall or look back through recent past events. In [20], a user study with a patient suffering from amnesia is conducted with SenseCam images and highlights the usefulness of these images in reminiscing about recent events by the patient. In [17], evidence is

found that SenseCam images do facilitate people's ability to connect to their recent past. Similar applications of turning lifelogging into a short-term memory aid can also be found in [21], and [22]. Another good example of this is the work by Browne et al. [23] who used the visual lifelog from a SenseCam to stimulate autobiographical recollection, promoting consolidation and retrieval of memories for significant events. All these clinical explorations seem to agree that visual lifelogging provides a "powerful boost to autobiographical recall, with secondary benefits for quality of life" [23, 3].

Diet monitoring: Diet monitoring is another application of visual lifelogging for medical purposes. Though dietary patterns have been proved as a critical contributing factor to many chronic diseases [24], traditional strategies based on self-reported information do not fulfill the task of accurate diet reporting. More usable and accurate ways to analyze dietary information about an individual's daily food intake are badly needed. Visual media like images and videos provide hugely increased sources of sensory observations about human activities among which food intake can be monitored for diet analysis. The application of visual lifelogging in diet monitoring can support both patients with obesity and health care professionals analysing diets. DietSense [24] is an example of such a lifelogging software system using mobile devices to support automatic multimedia documentation of dietary choices. The captured images can be *post facto* audited by users and researchers with easy authoring and dissemination of data collection protocols [24]. Professional researchers can also benefit in performing diet intake studies with the help of lifelog browsing and annotation tools. Both audio recorders and cameras are combined in [25]. According to [26], individuals self-reported energy intake frequently and substantially underestimates true energy intake, while Microsoft SenseCam wearable camera can help more accurately report dietary intake within various sporting populations.

Disease diagnosis: Project IMMED [27] is a typical application of lifelogging to ADL, the goal of which is assessing the cognitive decline caused by dementia. Audio and video data of the instrumented activities of a patient are both recorded in [27] and indexed for medical specialists' later analysis. In [28], a wearable camera is used to capture videos of patients' activities of daily living. A method for indexing human activities is presented for studies of progression of the dementia diseases. The indexes can then be used for doctors to navigate throughout the individual video recordings in order to find early signs of the dementia in everyday activities. The same rationale is also reported in [29]. Most recently, by combining biosensor information with frequent medical measurements, wearable devices proved to be useful in identification of early signs of Lyme disease and inflammatory responses [30].

ADL analysis: More concerns is now being shown in modern society about the individual health and well-being of everyday life. However, any long-term investigation into daily life comes across lots of difficulties in both research and the medical treatment area. Occupational therapy aims to analyze the correlation between time

spent and our actual health, and there is a growing body of evidence indicating the relationship [31, 32]. Observational assessment tools are needed to correctly establish care needs and identify potential risks. Long-term daily routines and activity engagement assessments are necessary to evaluate the impact on activities of daily living caused by diseases or old age, hence to provide a proper programme towards the needs of each patient. While traditional self-reporting or observational measures are time-consuming and have limited granularity, lifelogging can provide an efficient approach to providing broader insights into activity engagement. [33] has shown that wearable cameras represent the best objective method currently available to categorise the social and environmental context of accelerometer-defined episodes of activity in free-living conditions. In [34], the feasibility of using visual wearable cameras to reconstruct time use through image prompted interview is demonstrated.

2. Semantic Indexing of Visual Lifelogs: a Static View

While this chapter focuses on computer vision for lifelogging applications, this takes place within the context of a huge growth in the amount, and use, of generated multimedia content. Nowadays we are not just consumers of multimedia through the traditional channels of broadcast TV, movies, etc., we are also generators of multimedia, especially visual multimedia, though the widespread availability of smartphones and the strong support for sharing of our images and vidoes through our online social networks like Facebook and Twitter. While the ubiquitous mobile smartphone device has provided access to technology for creating and sharing of visual media and has catalysed the growth in such media creation so that we can easily create permanent memory records of our lives, this then creates the huge challenges to analyse, index, and retrieve from this visual media.

Perhaps the easiest way in which we provide access to visual media, from whatever source, is to use automatically created metadata. In fact this is true of all retrievable artifacts, whether analog, digitised or born digital, and to support that there is a legacy of decades of work in developing standards for metadata creation and access. This varies from Dublin Core [35] which is general purpose, to EXIF metadata for images taken with almost all cameras. From such simple metadata as date, time and location we can actually go quite far and support access to images based on grouping images from the same, or different, users into "events", and we can then augment the description of images and videos for these events using external data. For the specific case of lifelogging, there is past work which showcases such lifelog event augmentation [36] using tags from similar images taken at the same location and the same, or different times.

While this is inventive and satisfies some information seeking needs, and is completely automated, by adding even a small amount of *content* processing we can go

much further. Consider the image shown in Figure 1.1. Knowing the date, time and GPS coordinates from EXIF metadata, resolving this using a gazeteer, drawing some further linked data from Wikipedia or another open source, analysing the content to perform face detection and recognition against a database of friends' faces and resolving who the person is as well as detecting some simple setting characteristics, we can tag this image with the following . . . *1 person, date = "22 Feb 2007", time = " 2pm", setting = "outdoors, daylight", location = " Auron, Southern France", weather = "sunny", setting = "ski resort", altitude = " 1,622m", setting = "snow", setting = "manmade environment".*

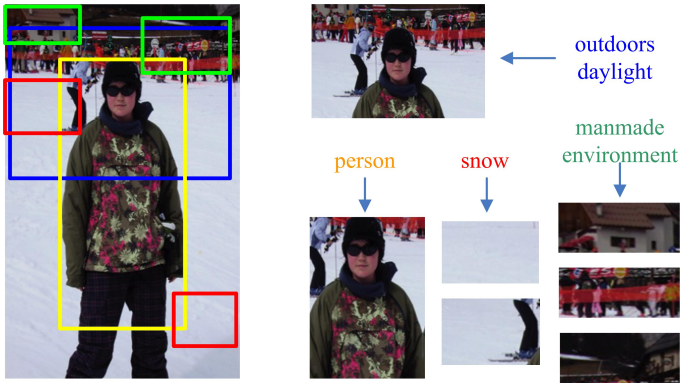

Figure 1.1 Using metadata and lightweight content analysis to tag an image.

The next obvious enhancement to describing image content is manual annotation. Annotation is the process of generating high level semantic metadata to describe something and has become the *de facto* norm in applications like Twitter where we use hashtags, in annotating the presence of our friends' faces in Facebook images, and there are many other cases where we use dedicated forms to collect structured data, like completing a web form for a quotation for car insurance. This is both boring, and not scalable, and even where we outsource a task like image annotation to a crowd as in Amazon's Mechanical Turk or microworkers, or even if we gamify the annotation through micro-games for online players [37] or more sophisticated games with a purpose [38],this still does not scale upwards to very large numbers plus there is a latency between time of image capture and time of annotation which could be important if we want to do anything in real time with a lifelog.

In reality, the only way to achieve real time analysis and content indexing of lifelog imagery is to automate the process and initial efforts in this area were to leverage the developments in semantic annotation of images, by pre-defined tags. Automatic an-

notation of images has been a long-standing application of machine learning and for many years researchers tried to use low-level image (and video) features as a basis for building classifiers for individual content tags. These were based on extracting features such as colour histograms, dominant colours, textures, and shapes from entire images and also from localised areas within images. In recent years additional features such as SIFT/SURF characteristics have been included and indeed the MPEG-7 standard which was explicitly defined to encode image and video metadata, supported this. Extracting low level features form an image is computationally fast and can be used to build a compact representation and hence a small feature space, for images. Once a ground truth of images annotated to the presence, or absence, of a semantic concept or tag is available, conventional machine learning tools can be used to learn the differences between those that have, and do not have, the particular tag, all done within the feature space of the low level features.

For many years, within the context of the TRECVid video benchmarking activity, researchers struggled to achieve high enough accuracy for the classifiers, as well as large enough numbers of tags in order to be usable [39]. Then, in 2012, things changed with the significant improvements in recognition accuracy obtainable when deep learning networks were applied to this computer vision problem for classifying and tagging images, all led by the work of Geoffrey Hinton's team [40]. Suddenly there was a perfect storm of conditions for effective and usable automatic image tagging — accuracy levels improved almost to human levels of agreement, large data repositories of tagged images became available through the manual annotation of user generated content, and the large computational requirements needed to train (and run) deep networks could be addressed by using relatively cheap GPUs.

The level of accuracy and the scale and size of these (independent) taggers is plateauing and levelling off and there are now emerging cloud services which can automatically tag. The output of one of these, `www.imagga.com` is shown in Figure 1.2. These services can easily be hosted on cloud services and we can expect them to be built into consumer photography and social media image processing. Indeed Google Photos and Facebook now tag uploaded images in this way, initially presented as an assistance to those with visual impairment.

While these are extremely useful, there is still a long way to go with this technology because, for example, semantic concepts are treated as independent of each other and there's no consolidation across the set of assigned tags (though as we show later in Section 3 we are making progress in this area).

When we examine the developments in semantic image tagging through the lens of lifelogging then for a lifelog, a simple set of semantic tags is not what we really want. This has recently been highlighted in work on the Kids'Cam project which annotated almost 1.5M images taken from wearable cameras by 169 pre-teen children in New Zealand in a project to measure children's exposure to fast food advertising [41]. Here

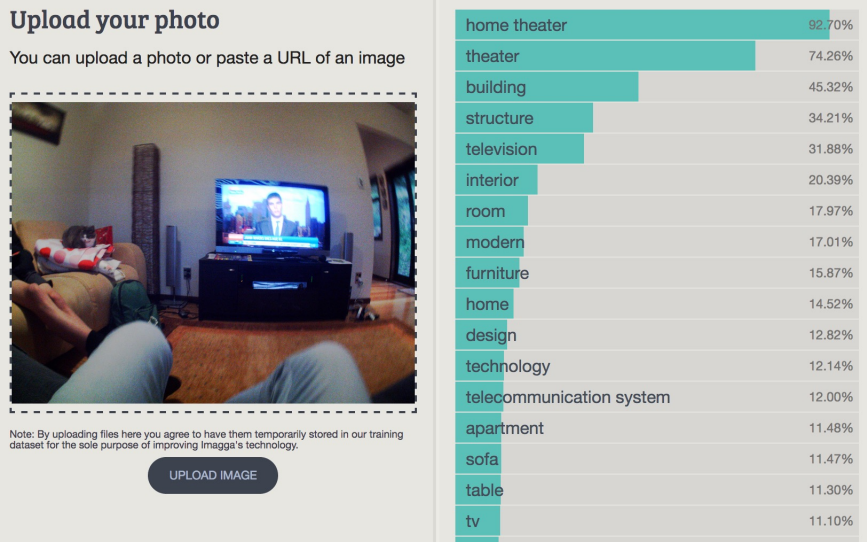

Figure 1.2 Automatically tagging a lifelog image using `www.imagga.com`.

we found that where there are errors in the machine learning recognition of semantic tags in a lifelog image, these need to be *pruned* in a much more dynamic manner from the set of tags.

3. Utilizing Contextual Semantics: a Dynamic View

In Section 2, the automatic indexing of visual lifelogs is discussed based the detection of concepts from low-level visual features. Though effectiveness have been shown using the state-of-the-art machine learning methods, current concept detectors are mostly the one-per-class classifiers which ignore the contextual correlation between concepts. However, the appearance of concepts is not independent with each other. Instead, the co-occurrence and re-occurrence patterns of various concepts implies that concepts interact during the evolution of everyday activity engagements. This section will elaborate the utilization of concept contextual semantics which benefits both the indexing of visual lifelogs and the characterization of everyday activities.

3.1. Modeling Global and Local Occurrence Patterns

One day's continuous visual lifelogs such as image streams can usually be segmented such as using the technique introduced in [42], into dozens of events which eases the representation and interpretation of everyday engagements. A lifelog event corresponds to a single activity in the wearer's day such as watching TV, commuting, or

eating a meal, and the durations vary a lot due to the engagement natures of various activities.

After applying automatic image indexing methods as introduced in Section 2, we can characterize each lifelog event as a series of concept detection results performed on each of the images representing it. Assume that event E_i consists of successive images $I^{(i)} = \{Im_1^{(i)}, Im_2^{(i)}, ..., Im_k^{(i)}\}$ and we obtained M automatic concept annotators in Section 2. By representing $Im_j^{(i)}$ with a vector of concept appearance $C_j^{(i)} = \{c_{j1}^{(i)}, c_{j2}^{(i)}...c_{jM}^{(i)}\}$, the total event set can be described as a confidence matrix $C_{N\times M}$, where $N = \sum_{i=1}^{n} k_i$ and k_i is the number of images in each event E_i.

Compared to low-level features like color, shape or texture features, the results of concept detection provides a more natural way to describe and index vision content which is close to human expectations. However, the initial detection results such as $C_{N\times M}$ is still noisy because the current machine learning methods are far from being perfect due to the dependence on the large volume of training corpora. Though breakthroughs have been achieved using deep learning [40], the effective transferring of learned models to different application domains such as visual lifelogs is still questionable.

Detection refinement or adjustment methods [43, 44, 45, 46, 47] represent a stream of post-processing method which can enhance detection scores obtained from individual detectors, allowing independent and specialized classification techniques to be leveraged for each concept. In this section, we introduce an approach which can exploit the inter-concept relationships implicity from concept detection results of $C_{N\times M}$ in order to provide better quality semantic indexing. This is in contrast to current refinement methods which learn inter-concept relationships explicitly from training corpora and then apply these to test sets. Because acceptable detection results can be obtained for concept with enough training samples, as witnessed by TRECVid benchmark [39] and ImageNet competition [48], it is feasible to utilise detections with high accuracies to enhance overall multi-concept detections since the concepts are highly correlated.

3.1.1. Factorizing Indexing Results

The framework of result factorization is to exploit the global pattern in concept occurrence context. The intuition behind the factorization method is that, the high-probable correct detection results are selected to construct an incomplete but more reliable matrix which can then be completed by a factorization method. Depending on the organization of concept detection result C, we can apply different forms of factorizations such as matrix or tensor factorizations in order to to overlay a consistency on the underlying contextual pattern of concept occurrences. Co-occurrence and re-occurrence patterns for concepts are a reflection of the contextual semantics of concepts since everyday concepts usually co-occur within images rather than in isolation. In some

potentially long-duration activities like "using computer", "driving", etc., the indicating concepts may appear frequently and repeatedly in the first-person view.

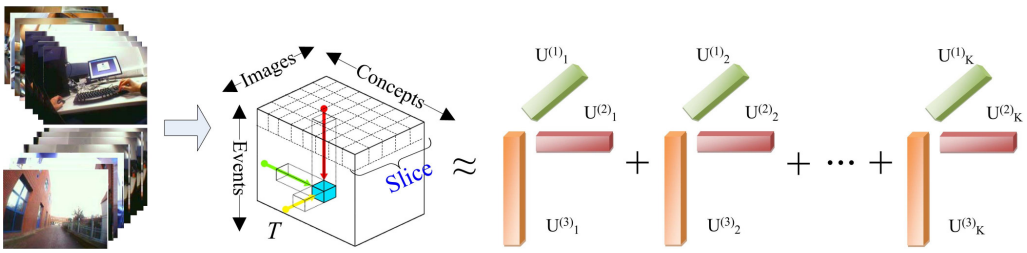

Figure 1.3 NTF-based concept detection enhancement framework.

As we can see from Fig. 1.3, the concept detection result C can be further represented as a 3-way tensor T. The rationale behind this is to avoid information loss from event segmentation and utilise the temporal features reflected in different events. As illustrated in Fig. 1.3, a tensor is advantageous in representing the structure of multi-dimensional data more naturally. The procedure for concept tensor construction and factorization is shown in Fig. 1.3. As illustrated, each slice is a segmented part of an event and is represented by a confidence matrix. The slices are then stacked one below another to construct a three-dimensional tensor which preserves the two-dimensional characters of each segment while keeping temporal features along the event dimension and avoids significant loss of contextual information.

In tensor factorization, the Canonical Decomposition (CD) [49] model simplifies the approximation of tensor C as a sum of 3-fold outer-products with rank-K decomposition $\hat{T} = \sum_{f=1}^{K} U_{\cdot f}^{(1)} \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)}$, which means each element $\hat{T}_{ijk} = \sum_{f=1}^{K} U_{if}^{(1)} U_{jf}^{(2)} U_{kf}^{(3)}$. This form of factorization constrains that each factor matrix has the same number of columns, i.e., the length of latent features has the fixed value of K. When $n = 2$, T is simply a 2-mode tensor which is indeed the initial matrix C and the factorization degenerates as $\hat{T} = \sum_{f=1}^{K} U_{\cdot f}^{(1)} \otimes U_{\cdot f}^{(2)} = U^{(1)} U^{(2)T}$.

The CD approximation factorization defined above can be solved by optimizing the cost function defined to quantify the quality of the approximation. For an arbitrary n-order tensor, the cumulative approximation error can be used to define the cost function, which has the form $F = \frac{1}{2}\|T - \hat{T}\|_F^2 = \frac{1}{2}\|T - \sum_{f=1}^{K} \otimes_{i=1}^{n} U_{\cdot f}^{(i)}\|_F^2$. In factorizing the confidence tensor, the weighted measure is more suitable since detection performance is different due to the characteristics of concepts and quality of the training set. Because each value c_{ij} in C denotes the probability of the occurrence of concept v_j in sample s_i, the estimation of the existence of v_j is more likely to be correct when c_{ij} is high, which is also adopted by [43, 50] under the same assumption that the initial

detectors are reasonably reliable if the returned confidences are larger than a threshold. To distinguish the contribution of different concept detectors to the cost function, the weighted cost function is employed as

$$\begin{aligned} F &= \frac{1}{2}\|T-\hat{T}\|_W^2 = \frac{1}{2}\|\sqrt[2]{W}\circ(T-\hat{T})\|_F^2 \\ &= \frac{1}{2}\sum_{i_1,i_2,...,i_n} w_{i_1,i_2,...,i_n}(T_{i_1,i_2,...,i_n} - \sum_{f=1}^{K}\otimes_{i=1}^{n} U_{\cdot f}^{(i)})^2 \\ &\text{s.t. } U^{(1)}, U^{(2)}, ...U^{(n)} \geq 0 \end{aligned} \tag{1.1}$$

where $\circ$ denotes element-wise multiplication, order-n (n-dimensional) tensor W and $T \in \mathcal{R}^{I_1\times I_2\times ...\times I_n}$ ($I_1, I_2, ..., I_n$ denotes the size of each of the tensor's dimensions), W denotes the weight tensor whose elements are larger for reliable and lower for less reliable detections, and $\|\cdot\|_F^2$ denotes the Frobenius norm, i.e., the sum of squares of all entries in tensor. The nonnegative constraints guarantees each component described by $U^{(i)}$ are additively combined. The discussion of the solution of the above formalized problem is out of the scope of this chapter, we provide the iterative updating rule using multiplicative method [51, 52], as:

$$U_{i_t f}^{(t)} \leftarrow U_{i_t f}^{(t)} \frac{\sum_{I-i_t}(W\circ T)_{i_t}\prod_{r\neq t} U_{i_r f}^{(r)}}{\sum_{I-i_t}(W\circ \hat{T})_{i_t}\prod_{r\neq t} U_{i_r f}^{(r)}}, 1 \leq t \leq n \tag{1.2}$$

where $I = i_1, i_2, ..., i_n$ is an n-tuple index whose value is in range of $i_t \in [1, I_t], 1 \leq t \leq n$, $I - i_t$ denotes the $n-1$ subset with i_t removed from I and $(\cdot)_{i_t}$ denotes the t-th dimension is assigned with value i_t for a given tensor. Taking 3-way tensor as an instance, the refinement can be expressed as a fusion of confidence tensors after factorization:

$$T' = \alpha T + (1-\alpha)\hat{T} = \alpha T + (1-\alpha)\sum_{f=1}^{K} U_{\cdot f}^{(1)} \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)} \tag{1.3}$$

3.1.2. Temporal Neighbourhood-Based Propagation

For much of the visual media we use for assitive living purpose there is a temporal aspect. For example the image streams captured continuously in visual lifelogging is inherently temporal as it captures imagery over time and thus they may have related content because they are taken from the same scene or have the same characters of related activities. For such "connected" visual media it makes sense to try to exploit any temporal relationships when post-processing initial concept detection, and to use the "neighbourhood" aspect of visual media.

Following refinement based on global context using high-order tensor or the degenerated matrix factorization, detection results will have been adjusted in a way consis-

tent with the latent factors modeled in the factorization. While this procedure exploits general contextual patterns which are modeled globally by factorization, the similarity propagation method can further refine the result by exploiting any local relationships between samples.

A new confidence matrix C' can be recovered from the refined tensor T' and C' has the rows and columns representing image samples and concepts respectively. As a refined result of $C_{N\times M}$, C' can provide better measures to localize highly related neighbours for similarity-based propagation. The similarity between samples s_i and s_j can be calculated by Pearson Correlation, formulized as:

$$P_{i,j} = \frac{\sum_{k=1}^{M}(c'_{ik} - \bar{c'}_i)(c'_{jk} - \bar{c'}_j)}{\sqrt{\sum_{k=1}^{M}(c'_{ik} - \bar{c'}_i)^2}\sqrt{\sum_{k=1}^{M}(c'_{jk} - \bar{c'}_j)^2}}$$

where $c'_i = (c'_{ik})_{1\le k\le M}$ is the i-th row of C', and $\bar{c'}_i$ is the average weight for c'_i. To normalize the similarity, we employ the Gaussian formula and denote the similarity as $P'_{i,j} = e^{-(1-P_{i,j})^2/2\delta^2}$, where δ is a scaling parameter for sample-wise distance. Based on this we can localize the k nearest neighbours of any target sample s_i.

The localized k nearest neighbours can be connected with the target sample using an undirected graph for further propagation. For this purpose, the label propagation algorithm [53] is derived to predict more accurate concept detection results based on this fully connected graph whose edge weights are calculated by the similarity metric as calculated by P'_{ij}. Mathematically, this graph can be represented with a sample-wise similarity matrix as $G = (P'_{i,j})_{(k+1)\times(k+1)}$, where the first k rows and columns stand for the k nearest neighbours of a target sample to be refined which is denoted as the last row and column in the matrix. The propagation probability matrix G' is then constructed by normalizing G at each column as

$$g'_{i,j} = \frac{P'_{i,j}}{\sum_{l=1}^{k+1} P'_{l,j}}$$

which guarantees the probability interpretation at columns of G'. By denoting the row index of k nearest neighbours of a sample s_i to be refined as n_i $(1 \le i \le k)$ in C' and stacking the corresponding rows one below another, the neighbourhood confidence matrix can be constructed as $C_n = (c'_{n_1}; c'_{n_2}; ...; c'_{n_k}; c'_i)$. The propagation algorithm is carried out iteratively by updating

$$C_n^t \leftarrow G' C_n^{t-1} \tag{1.4}$$

where the first k rows in C_n stand for the k neighbourhood samples in C' indexed by subscript n_i and the last row corresponds to the confidence vector of the target sample s_i. Since C_n is a subset of C', the graph G constructed on C_n is indeed a subgraph of

the global graph constructed on C'. During each iteration, the neighbourhood concept vector c'_{n_i} needs to be clamped to avoid fading away. After a number of iterations, the algorithm converges to a solution in which the last row of C_n is a prediction based on similarity propagation. In this way, the local relationships between neighbours can be used for a more comprehensive refinement.

3.2. Attribute-Based Everyday Activity Recognition

Though concept detection and refinement can provide a semantic representation for visual assistive living, many assistive living applications require the capability to recognize semantic concepts which have a temporal aspect corresponding to activities or events, i.e., characterizing the whole time series rather than merely interpreting single images. While low-level feature-based methods have been shown to be ill-suited for multimedia semantic indexing due to the lack of semantics for user interpretation, high dimensionality, etc., high-level concept attributes are widely employed in the analysis of complex semantics corresponding to things like events and activities. Since such semantic structures can be represented as typical time series, the recognition of events or activities can be regarded as dynamics-based recognition using concept detection results. This is usually carried out by representing the time series as a sequence of units such as video clips or image frames. After concatenating the results of concept detectors on each unit, time series can then be represented by a temporally-ordered sequence of vectors, as shown in Fig. 1.4.

Fig. 1.4 shows the paradigm of utilising concept temporal dynamics for high-level activity detection, in which typical indoor activities like "cooking", "watching TV", etc. are demonstrated, as well as the corresponding trajectories. The concept detection results temporally aligned with these activities are depicted as confidence bars in the diagram, to represent the likelihood of the presence and absence of concepts, respectively. It is important to note that the concept detection shown in Fig. 1.4 does have errors and this can affect further analysis to various degrees. Therefore, in order not to propagate these errors into the subsequent analysis for activity and behaviour characterization, the original concept detections can be enhanced using the refinement methods as introduced in Section 3.1.

As shown in Fig. 1.4, everyday activities can be regarded as stochastic temporal processes consisting of various lengths of concept vectors. With this, the dynamic evolution of concept vector occurrences can characterize a deeper meaning of underlying, or derived, human activities if the evolution patterns can be modeled. Attribute-based event and activity detection has attracted much research attention. More importantly, it is found that although state-of-the-art concept detections are far from perfect, they still provide useful clues for event classification [54]. [55] also revealed that this representation outperforms – and is complementary to – other low-level visual descriptors

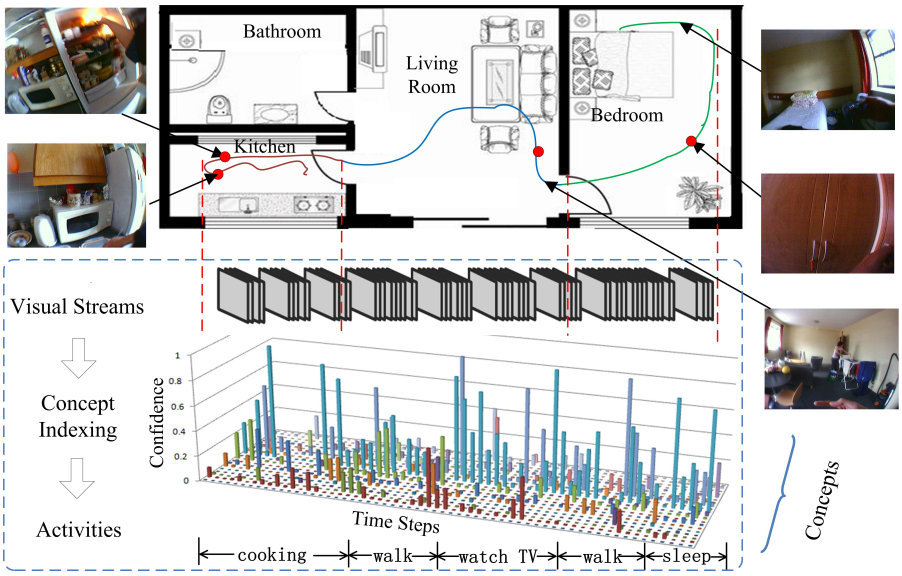

Figure 1.4 The dynamics of concept attributes quantified by confidences returned by concept detections.

Methods	Temp?	#Types	#Concepts
Max-Pooling	×	10 [58], 15 [59], 25 [60]	50 [58], 101 [59], 93 [60]
Bag-of-Features	×	10 [58], 18 [61], 3 [55]	50 [58], 42 [61], 280 [55]
Temporal Pyramids	√	18 [61]	42 [61]
Graph Model	√	16 [3, 62]	85 [3, 62]
Fisher Vector	√	16 [3], 15 [63]	85 [3], 60 [63]
Dynamic System	√	25 [60], 5 [64]	93([60], [64])

Table 1.1 Some typical time series recognition methods.

for event modeling. This is also tested in a more comprehensive visual lifelogging experiment in [56]. Similar work is also carried out using concept detections to characterize everyday activities as reported in [57, 3] where activity recognition is built on the basis of concept detection. In this chapter, we investigate a selection of attribute-based recognition methods which are suitable for activity characterization in assistive living. These time series recognition methods investigated are summarized in Table 1.1 [56]. Whether they utilise temporal features, the number of event/activity types and the number of concept attributes are all depicted in Table 1.1. The details of the different recognition methods we implemented are now outlined.

Max-Pooling (MP): As one of the fusion operations for concept detection results,

Max-Pooling [58] has been demonstrated to give better performance compared to other fusions for most complex events. In Max-Pooling, the maximum confidence is chosen from all keyframe images (or video subclips) for each concept to generate a fixed-dimensional vector for an event or activity sample. Since by definition the maximum value cannot characterize a temporal evolution of concepts within a time series, this method can be regarded as non-temporal.

Bag-of-Features (BoF): Similar to Max-pooling, Bag of Features is a way of aggregating concept detection results by averaging the confidences over time window. Because Bag-of-Features and Max-Pooling reflect the statistical features within the holistic time series, they both ignore the temporal evolution of concept detection results.

Temporal Pyramids (TP): Motivated by the spacial pyramid method, the temporal pyramids proposed in [61] approximate temporal correspondence with a temporally-binned model [65]. In this method, the histogram over the whole time series represents the top level while the next level can be represented by concatenating two histograms of temporally segmented sequences. More fine-grained representations can be formalized in the same manner. By applying the multi-scale pyramid to approximate a coarse-to-fine temporal correspondence, this method generates fixed-length features with temporal embeddings.

Graph Model: A generative method based on Hidden Markov Model (HMM) as used in [3] is employed in this representation. HMMs are first trained for each activity class and the log-likelihood representations of per-class posteriors can be concatenated into a vector. While HMM assumes all of the observations in a sequence are independent and conditional on the hidden states, Conditional Random Field (CRF) has no such constraints so that it allows the existence of non-local dependencies between hidden states and observations. Because the dependence is allowed in a wider range, a CRF model can flexibly adapt to dynamic sequences in which high correlations might exist between different regions. This is especially useful for modeling the kind of activities recorded by wearable lifelog cameras. In [62], the hidden CRF is applied to characterize everyday activities.

Fisher Vector (FV): The principle of the Fisher kernel is that similar samples should have similar dependence on the generative model, i.e. the gradients of the parameters [66]. Instead of directly using the output of generative models, such as in the HMM method, using a Fisher kernel tries to generate a feature vector which describes how the parameters of the activity model should be modified in order to adapt to different samples. Concept observations X can be characterized as Fisher scores with regard to the parameters λ, $U_X = \nabla_\lambda logP(X|\lambda)$. Therefore, the Fisher kernel can be formalized as $K(X_i, X_j) = U_{X_i}^T I_F U_{X_j}^T$, where $I_F = E_X(U_X U_X^T)$ denotes the Fisher information matrix.

Liner Dynamic System (LDS): As a natural way of modeling temporal interaction

within time series, Liner Dynamic Systems [60] can characterize temporal structure with attributes extracted from within a sliding window. The time series can be arranged in a block Hankel matrix H whose elements in a column have the length of sliding window (denoted as r) and successive columns are shifted with one time step. According to [60], singular value decomposition of $H \cdot H^T$ has achieved comparable performance to more complex representations. We constructed the feature using the k largest singular values along with their corresponding vectors.

The application of above discussed methods in characterizing everyday activities is based on the assumption that the concepts contained in these images can reflect significant temporal patterns, characterizing the semantics of the activities they represent. In this case, the temporal consistency of certain types of concepts like "indoor", "screen" and "hands" can be viewed as cues for concepts like "using computer". Even though some activities require the user to be changing their location all the time like "walking" and "doing housework", the dynamics of concepts present in the activity will still show some patterns, such as the frequent appearance of "road" for the "walking" activity in an urban environment, or the transitions between "kitchen" and "bathroom" for the "housework" activity. Therefore, appropriate concept selection is necessary to characterize the dynamic evolution of time series based concepts. According to [56, 67], the recognition performances of activities can be enhanced accordingly if more appropriate concepts are utilized. This is more obvious when the original concept detections are less satisfactory [56]. In practice, the assistive living developers who want to apply attribute-based activity recognition should focus on three affecting factors including concept selection, concept detection and activity recognition methods.

4. Interacting with Visual Lifelogs

When considering the various approaches that have been taken to lifelog data visualisation and user interfaces, it requires an understanding of the application use-cases and types of data involved. In a simple implementation of assistive living personal data access, and one that many readers will be familiar with, the personal quantified self data gathered by wearable fitness trackers and related sensors is typically aggregated into meaningful infographics and charts, summarising the activities or performance of the individual across a number of useful dimensions. However, given our definition of lifelogging that includes wearable camera data, along with potentially other sources of data, then this simple data aggregation summarisation approach will not be appropriate and more advanced forms of visualisation are required [5].

There has been some initial work on design considerations for lifelog content. Whittaker et al. [68] proposed a set of design principles for developing lifelogging systems that support human memory and Hopfgartner et al. [69] presented a set of

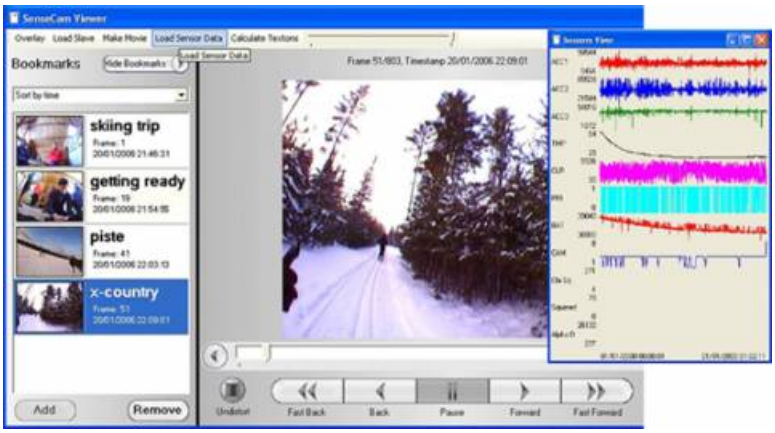

Figure 1.5 A rapid-playback Lifelog Interface

user interaction templates for the design of lifelogging system. Additionally, Byrne et al. [70] presented a set of guidelines for the presentation and visualisation of lifelog content, including the need for a simple and intuitive interface, segmenting the content into comprehensible units, aiding human memory and support for exploration and comparison. All of these offer meaningful insights into how to design the user interface and user experience for lifelog applications.

However, the first applications of wearable camera data for assistive living were concerned with utilising lifelog technologies to support memory studies of the individual. These studies have been concerned with supporting the individual to recollect past experiences, either as a form of scientific experimentation or as an assistive technology for individuals with memory impairments. Van den Hoven et al. [71] provide a set of design considerations that support using lifelogs to help people remember in everyday situations. These are based around the concept that when recollecting, reliving past experiences, we know that visual media, especially captured from the first person viewpoint, provide very powerful memory cues and lead to what is referred to as Proustian moments of recall [72]. Hence the early lifelog interaction scenarios to support human memory have focused on assisting the user to browse through lifelog archives. The initial work in the area from Microsoft Research presented the SenseCam image browser which facilitated rapid playback of image sequences, along with a manual event segmentation tool [23]. Such an interface presents a stop-motion style rapid playback through visual life logs and they provide us with a baseline visualisation of visual lifelog data, though one that quickly presents challenges to the user due to the lack of data organisation when the archive of data grows to weeks, months or even years [5].

A first effort at organising lifelog data was the event-based SenseCam image browser

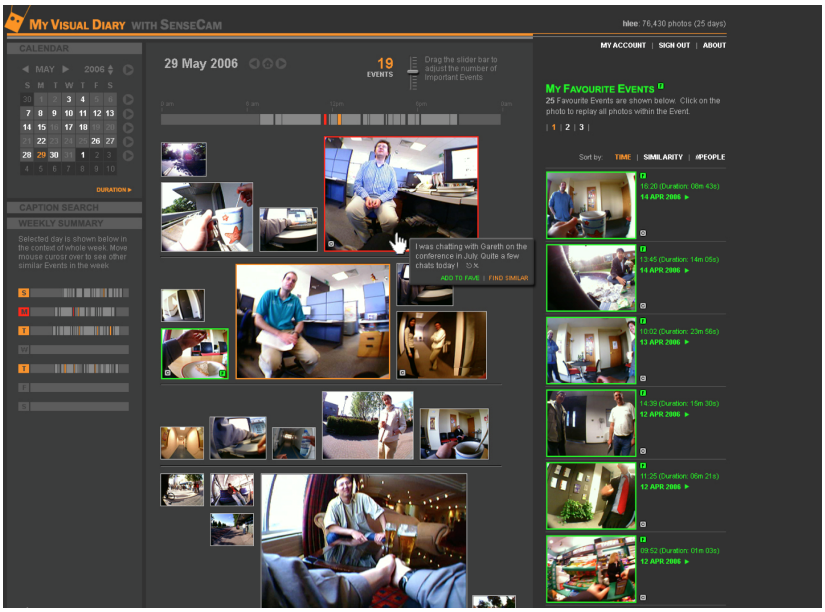

Figure 1.6 An event-based Lifelog Browser

[73] which automatically segmented lifelog data into events and then allowed users to manually annotate those events, thereby supporting a basic level of structured browsing and linkage of lifelog data. The key organisational metaphor for the data here was as a diary with a chronological ordering of events on any given day. It is worth pointing out that an event-level browser or lifelog data can reduce many thousands of images per day to a more manageable set of events (typically about 30-40). Such interfaces usually include drill-down functionality to enable the user to examiner some or all images that occur within one selected event. An example of an event-level browser [73] is shown in Fig. 1.6. Another related approach was ShareDay [74] which allows both browsing through lifelog events, and event sharing with family and friends.

One of the key challenges facing any life logger or lifelog practitioner is the volume of data to be organised. Event browsers are a step in the right direction, in that they convert a temporal stream of lifelog data into a chronologically ordered sequence of events. However, even with segmenting a lifelog into events, it is still challenging for the lifelogger or practitioner to actually find anything in a large archive. We know from prior research into personal photograph collections that a rich annotation mechanism enhances user access [75]. The inclusions of annotations, at the image or event level, supports a user in searching through lifelog archives. One example is the food search and browse engine from Kitamura et al. [76], that encourages users to manually annotate food events within a lifelog archive. However relying on human

annotations is inherently problematic as users are unlikely to expend a lot of effort in manually tagging large volumes of lifeblood data. Hence, it is necessary to incorporate automatic annotation of lifelog data, using approaches such as these outlined in this chapter. In this way, users can expect to be able to perform automated search through the lifelog archives and data abstraction and linkage models could be developed to support user access. Early work in this area includes the ethnography browser which automatically annotates lifelog events for lifestyle traits and real-world logos and objects [77], which was developed as a proof-of-concept lifelog search engine allowing cross-demographic analytics of lifelogs.

While there has been some effort in developing richer annotation tools and even in developing automated search tools for lifelog data ([78], [79], [80], [81]), there is no standard approach and progress on interactive lifelog interfaces has been even slower. In 2015, the first collaborative effort to develop high-quality search engines, some with visual interfaces [82] have been developed as part of the NTCIR-Lifelog collaborative benchmarking exercise[83]. Over the coming years, initiatives such as NTCIR-Lifelog and ImageClef-Lifelog are expected to expedite the progress in developing new interfaces and visualisations of lifelog data.

There have also been a number of interfaces developed that specifically provide assistive technologies to assist users in visualising, comprehending and accessing lifelog data. Caprani et al. [84] developed a touchscreen browser that incorporated event segmentation to allow computer illiterate older adults to browse and share lifelog data. Gurrin et al. developed a gesture-based event browser for use on the livingroom TV [85] which utilised the Nintendo Wii platform to browse through lifelog data via a gesture-based interface. Finally a colour-based data abstraction [86] in which temporal lifelog visual data was represented by a colour-of-life wheel on the desktop and touchscreen devices.

These represent a summary of the different user interfaces and access methodologies that have been developed to search and browse through (mostly visual) lifelog data, however there have been no clear evaluation of the effectiveness of these approaches in comparison to each other. The best solution may represent a combination of some of the aforementioned approaches, but this requires future research. Initiatives such as NTCIR-Lifelog are providing the impetus to encourage progress in this area. Extending the existing annotation approaches to incorporate user activities and visual concepts will provide a richer annotation of the data, resulting in the assumption of better quality search and browsing support for lifelog data. However, this is still work in progress, and it is likely to be a number of years before lifelogs get their equivalent of the top ten ranked list in web search.

5. Conclusion and Future Issues

While the sensing facilities are relative mature thanks to the quick emergence of smart mobiles and wearable devices which are lightweight and embedded with powerful sensing capabilities and large storage, the understanding of visual lifelogging for the purpose of assistive living is in its infancy due to the challenges induced by wearers' movement, visual diversity, etc. Meanwhile, how to design novel and effective interactive applications is another important factor for changing visual lifelogging from a write-only memory surrogates. These two aspects of understanding and interactive design of visual lifelogging are discussed in above sections in order to make visual lifelogging more adoptable in assistive living, including the understanding of concepts and activities, the visualization and interactions for possible lifelogging applications, which can be summarized as:

- **Understanding of visual lifelogging**: As previously discussed in this chapter, the semantic indexing of visual lifelogging is based on an automatic mapping from visual features to a set of concepts. By utilising the contextual correlations of concepts, the refinement of concept indexing results can help to improve the understanding of visual lifelogging by effectively modelling the global and local occurrence patterns. Finally, a segment of activity can be recognized by characterizing the time series representation with attribute-based everyday activity recognition.
- **Interactive visual lifelogging**: We have seen that preliminary research has been carried out on the development of access methodologies for visual lifelogging, yet there still remains significant challenges to overcome. For example, there is not yet a clear understanding of how best to present visual lifelogs on a user interface; it is likely that some query-dependent visualisation and access methodology would be required. Exactly how to develop such lifelog search engines is also unknown. What we can state is that initiatives such as NTCIR-Lifelog are supporting the first comparative efforts at the development of visual lifelog search engines and analytics interfaces.

Though preliminary research have been carried out towards the understanding and interaction of visual lifelogging which can benefit assistive living in various aspects. Visual lifelogging is still faced with challenges when satisfying the needs of assistive living. Some future directions can be summarized as:

- **Real-time concerns**: The recognition accuracy is currently the main research concern in understanding everyday activities in assistive visual lifelogging. While some analysis can be carried out offline like life pattern analysis, diet monitoring, memory aid, etc., some other applications for assistive living requires further adoption of such technical schemes with real-time undersanding such as context-aware personalized service, smart home, behaviour monitoring, etc. Similar as humans in learning about the contents in the image at a glance, the reframing of object detec-

tion can be performed more efficiently in a deep neural network [87] with recent progresses. Another possible solution to tackle the efficiency in understanding assistive visual lifelogging is the distributed data processing strategy. For example, in [88, 89] the elastic nature of cloud infrastructure can be utilized to keep a tradeoff between performance and cost. In this case, the server side responsible for high computational complexity processing is an enabler for quick recognition of everyday concepts and activities. In addition, the fusion of heterogeneous senor readings can also tackle the scalability and performance challenges of real-time analysis in assistive living, i.e. substituting the high power/memory consuming sensor by a lower one when the lower power/memory consuming sensor data can be fused and maintain comparable activity recognition accuracy [90].

- **Incorporating with external knowledge**: Compared to some other forms of visual media which can be collected from the Internet publicly, it is more difficult to gather large collections of visual lifelogs for research or experimental purposes, as this requires enough volunteers to continuously collect their everyday activities. In addition, due to the concerns of privacy issues, making such collections to others are unrealistic so far. Transfer learning [91] is a useful method in improving the learning of the target predictive function using the knowledge learned from a source domain. Transfer learning is particularly valuable when the number of labels in target domain is much smaller than that in source domain. How to effectively transfer the learned knowledge from large volume datasets such as ImageNet [48] to the characterization of everyday activities for assistive living is definitely an research direction to be exploited. As introduced in Section 3.2, appropriate selection of semantic concepts as attributes impacts the recognition accuracy of activities. Though manual construction of semantic space such as topic-related concept selection user experiment [92, 93], has shown its merits, such method is less flexible when dealing with various activity types which are not existing in the pre-defined activity set. A more feasible solution is to exploit external online resources in order to reflect human knowledge on activity-specific concept selection. [94] points out an interesting method by choosing WikiHow online forum to extract related tags from human daily life event queries. The transfering of hierarchical structure constructed in such way to concept-based representation of activity recognition can be another direction in visual lifelogging understanding.
- **Automatic captioning visual lifelog**: One of the reasons why a set of tags is insufficient as a representation of a lifelog is that a lifelog tells a story of a person's day, and a set of tags doesn't do this. Even a set of tags for a set of images, arranged into some sequence, doesn't tell a story. There are things in everybody's day that are routine, regular, and maybe they are important to be included in the story of the day. There are also unexpected things in our day, surprises, unanticipated events, and maybe they form part of the story too. So the most relevant

work on automatically describing what is appearing in visual lifelogs is automatic captioning [95], of images or of videos and it is only very recently that we are seeing work emerging from research groups which is showing this. The quality of this work is mostly poor, there are many problems to be overcome. For example, in image tagging we tend to tag objects that appear in the image, like bicycles and people and trees, rather than activities and actions like running or jumping or greeting somebody. Detecting activities and actions is much more complex than detecting objects, especially when the actions are spread over time, like "preparing a cup of tea" or "eating a meal". In 2016 the TRECVid activity introduced a task on automatic captioning of video [39] where the target videos to be captioned were social media videos from the Vine website, of up to 8 seconds each. For some videos, captioning was comparable to manual annotation but for many videos the automatic captions were poor because they lacked context, i.e. the information on the events and activities surrounding the video clip in question.

- **Ethical and privacy issues**: Because this book is about computer vision, the detailed discussion of these issues is out of the scope of this chapter. However, the social sensitivity on these issues related to new technologies and devices such as Google Glass are concerns to be taken while developing assistive lifelogging. Given current breakthroughs in computer vision and artificial intelligence, the ethical and privacy issues should no longer be the obstructive factors which might impede the development of computer vision in lifelogging. On the contrary, the maturity of assistive computer vision can help to deal with these issues more effectively. Privacy can be tackled by extracting features locally on mobile devices as the computational capability are improved, or through edge routers at home. The detection of privacy-related images or streams will be more accurate which can help to filter them out. In this case, both the privacy concerns of lifelogger and bystanders who might be captured by visual devices can be alleviated. Currently, we can deal with the privacy issues by integrate "privacy by design" into the development process [5, 96, 97], i.e., embedding the issues as core considerations throughout the entire life cycle of this technology. More specifically, seven principles [98] can be taken into account when developing a privacy-aware lifelogging systems. Similarly, some other ethical issues can be dealt with embedding necessary principles in the framework such as discussed in [99].

ACKNOWLEDGMENTS

This chapter is supported by Natural Science Foundation of China under Grant No. 61502264, 61210008, Beijing Key Laboratory of Networked Multimedia, and Science Foundation Ireland under grant number SFI/12/RC/2289.

BIBLIOGRAPHY

1. Allen AL, Dredging-up the past: Lifelogging, memory and surveillance. New Yorker 2007; :38–44.
2. Doherty AR, Caprani N, Conaire CO, Kalnikaite V, Gurrin C, Smeaton AF, et al., Passively recognising human activities through lifelogging. Computers in Human Behavior 2011; 27:1948–1958. URL: http://dx.doi.org/10.1016/j.chb.2011.05.002, doi:http://dx.doi.org/10.1016/j.chb.2011.05.002.
3. Wang P, Smeaton A, Using visual lifelogs to automatically characterise everyday activities. Information Sciences 2013; 230:147–161.
4. Bush V, As we may think. The Atlantic Monthly 1945; .
5. Gurrin C, Smeaton AF, Doherty AR, Lifelogging: Personal big data. Fundations and Trends in Information Retrieval 2014; 8:1–107. doi:10.1561/1500000033.
6. Wang DH, Kogashiwa M, Kira S, Development of a new instrument for evaluating individuals' dietary intakes. J Am Diet Assoc 2006; 106:1588–1593.
7. Alexy U, Sichert-Hellert W, Kersting M, Schultze-Pawlischko V, Pattern of long-term fat intake and bmi during childhood and adolescence-results of the donald study. Int J Obes Relat Metab Disord 2004; 28:1203–1209.
8. Placelab (MIT), http://web.mit.edu/cron/group/house_n/placelab.html. Last accessed: Mar. 2017.
9. Jalal A, Kamal S, Real-time life logging via a depth silhouette-based human activity recognition system for smart home services. In: Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on, 2014, pp. 74–80, doi:10.1109/AVSS.2014.6918647.
10. Jalal A, Kim Y, Kim D, Ridge body parts features for human pose estimation and recognition from rgb-d video data. In: Computing, Communication and Networking Technologies (ICCCNT), 2014 International Conference on, 2014, pp. 1–6, doi:10.1109/ICCCNT.2014.6963015.
11. Jalal A, Sharif N, Kim JT, Kim TS, Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart homes. Indoor and Built Environment 2013; 22:271–279.
12. Song Y, Tang J, Liu F, Yan S, Body surface context: A new robust feature for action recognition from depth videos. Circuits and Systems for Video Technology, IEEE Transactions on 2014; 24(6):952–964. doi:10.1109/TCSVT.2014.2302558.
13. Jalal A, Kamal S, Kim D, A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. Sensors 2014; 14(7):11735–11759.
14. Hori T, Aizawa K, Context-based video retrieval system for the life-log applications. In: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR '03, New York, NY, USA: ACM, 2003, pp. 31–38. ISBN 1-58113-778-8, doi:http://doi.acm.org/10.1145/973264.973270. URL: http://doi.acm.org/10.1145/973264.973270.
15. Mann S, Fung J, Aimone C, Sehgal A, Chen D, Designing EyeTap digital eyeglasses for continuous lifelong capture and sharing of personal experiences. In: Proc. CHI 2005 Conference on Computer Human Interaction, Portland, Oregon, USA: ACM Press, 2005.
16. Blum M, Pentland AS, Tröster G, InSense: Interest-based life logging. Multimedia, IEEE 2006; 13(4):40 –48. doi:10.1109/MMUL.2006.87.
17. Sellen A, Fogg A, Aitken M, Hodges S, Rother C, Wood K, Do life-logging technologies support memory for the past? An experimental study using SenseCam. In: Proc. CHI 2007, ACM Press, New York, NY, USA, 2007, pp. 81–90.
18. Gemmel J, Bell G, Lueder R, Mylifebits: a personal database for everything. Communications of the ACM 2006; 49(1):88–95.
19. Bell G, Gemmell J, Total Recall: How the E-Memory Revolution Will Change Everything. 2009.
20. Hodges S, Williams L, Berry E, Izadi S, Srinivasan J, Butler A, et al., SenseCam: A retrospective memory aid. In: Proc. 8th International Conference on Ubicomp, Orange County, CA, USA, 2006, pp. 177–193.

21. Berry E, Kapur N, Williams L, Hodges S, Watson P, Smyth G, et al., The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report. Neuropsychological Rehabilitation 2007; 17(4-5):582–601. URL: http://dx.doi.org/10.1080/09602010601029780, doi:10.1080/09602010601029780.
22. Vemuri S, Bender W, Next-generation personal memory aids. BT Technology Journal 2004; 22(4):125–138. doi:http://dx.doi.org/10.1023/B:BTTJ.0000047591.29175.89.
23. Browne G, Berry E, Kapur N, Hodges S, Smyth G, Watson P, et al., Sensecam improves memory for recent events and quality of life in a patient with memory retrieval difficulties. Memory 2011; 19(7):713–722. doi:10.1080/09658211.2011.614622.
24. Reddy S, Parker A, Hyman J, Burke J, Estrin D, Hansen M, Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype. In: EmNets'07: Proceedings of the 4th workshop on Embedded networked sensors, Cork, Ireland: ACM Press, 2007, pp. 13–17.
25. Kaczkowski CH, Jones PJH, Feng J, Bayley HS, Four-day multimedia diet records underestimate energy needs in middle-aged and elderly women as determined by doubly-labeled water. Journal of Nutrition 2000; 130(4):802–5.
26. O'Loughlin G, Cullen SJ, McGoldrick A, O'Connor S, Blain R, O'Malley S, et al., Using a wearable camera to increase the accuracy of dietary analysis. American Journal of Preventive Medicine 2013; 44(3):297 – 301. URL: http://www.sciencedirect.com/science/article/pii/S074937971200863X, doi:http://dx.doi.org/10.1016/j.amepre.2012.11.007.
27. Mégret R, Dovgalecs V, Wannous H, Karaman S, Benois-Pineau J, Khoury EE, et al., The IMMED project: wearable video monitoring of people with age dementia. In: Proceedings of the international conference on Multimedia, MM '10, New York, NY, USA: ACM, 2010, pp. 1299–1302. ISBN 978-1-60558-933-6, doi:http://doi.acm.org/10.1145/1873951.1874206. URL: http://doi.acm.org/10.1145/1873951.1874206.
28. Karaman S, Benois-Pineau J, Megret R, Pinquier J, Gaestel Y, Dartigues JF, Activities of daily living indexing by hierarchical HMM for dementia diagnostics. In: The 9th International Workshop on Content-Based Multimedia Indexing (CBMI), Madrid, Spain, 2011, pp. 79–84, doi:10.1109/CBMI.2011.5972524.
29. Mégret R, Szolgay D, Benois-Pineau J, Joly P, Pinquier J, Dartigues JF, et al., Wearable video monitoring of people with age dementia : Video indexing at the service of healthcare. In: International Workshop on Content-Based Multimedia Indexing, CBMI '08, London, UK, 2008, pp. 101–108.
30. Li X, Dunn J, Salins D, Zhou G, Zhou W, Schu SM, Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. PLoS Biology 2017; 15(1). doi:10.1371/journal.pbio.2001402.
31. Law M, Steinwender S, Leclair L, Occupation, health and well-being. Canadian Journal of Occupational Therapy 1998; 65(2):81–91.
32. McKenna K, Broome K, Liddle J, What older people do: Time use and exploring the link between role participation and life satisfaction in people aged 65 years and over. Australian Occupational Therapy Journal 2007; 54(4):273–284.
33. Doherty AR, Kelly P, Kerr J, Marshall S, Oliver M, Badland H, et al., Using wearable cameras to categorise type and context of accelerometer-identified episodes of physical activity. International Journal of Behavioral Nutrition and Physical Activity 2013; 10(1):22. URL: http://dx.doi.org/10.1186/1479-5868-10-22, doi:10.1186/1479-5868-10-22.
34. Developing a method to test the validity of 24 hour time use diaries using wearable cameras: A feasibility pilot. PLoS ONE ; 10, number =.
35. Weibel S, The dublin core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science and Technology 1997; 24(1):9–11.
36. Doherty AR, Smeaton AF, Automatically augmenting lifelog events using pervasively generated content from millions of people. Sensors 2010; 10(3):1423–1446.
37. von Ahn L, Dabbish L, Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, New York, NY, USA: ACM, 2004, pp. 319–326. ISBN 1-58113-702-8, doi:10.1145/985692.985733. URL: http://doi.acm.org/10.1145/985692.985733.

38. Von Ahn L, Dabbish L, Designing games with a purpose. Communications of the ACM 2008; 51(8):58–67.
39. Awad G, Snoek CG, Smeaton AF, Quénot G, Trecvid semantic indexing of video: A 6-year retrospective. ITE Transactions on Media Technology and Applications 2016; 4(3):187–208.
40. Krizhevsky A, Sutskever I, Hinton GE, Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, 2012, pp. 1097–1105.
41. Smeaton AF, McGuinness K, Gurrin C, Zhou J, OConnor NE, Wang P, et al., Semantic indexing of wearable camera images: Kids'cam concepts. In: Proceedings of the ACM Multimedia 2016 Workshop on Vision and Language Integration Meets Multimedia Fusion Amsterdam, The Netherlands October 16, 2016, ACM Press, 2016.
42. Lee H, Smeaton AF, O'Connor NE, Jones GJF, Blighe M, Byrne D, et al., Constructing a SenseCam visual diary as a media process. Multimedia Systems 2008; 14(6):341–349. URL: http://dx.doi.org/10.1007/s00530-008-0129-x, doi:10.1007/s00530-008-0129-x.
43. Kennedy LS, Chang SF, A reranking approach for context-based concept fusion in video indexing and retrieval. In: CIVR'07, 2007, pp. 333–340. ISBN 978-1-59593-733-9, doi:10.1145/1282280.1282331. URL: http://doi.acm.org/10.1145/1282280.1282331.
44. Wang C, Jing F, Zhang L, Zhang HJ, Image annotation refinement using random walk with restarts. In: ACM MM'06, 2006, pp. 647–650. ISBN 1-59593-447-2, doi:10.1145/1180639.1180774. URL: http://doi.acm.org/10.1145/1180639.1180774.
45. Wang C, Jing F, Zhang L, Zhang HJ, Content-based image annotation refinement. In: CVPR'07, 2007, pp. 1–8, doi:10.1109/CVPR.2007.383221.
46. Jiang YG, Wang J, Chang SF, Ngo CW, Domain adaptive semantic diffusion for large scale context-based video annotation. In: ICCV'09, 2009, pp. 1420–1427, doi:10.1109/ICCV.2009.5459295.
47. Jiang YG, Dai Q, Wang J, Ngo CW, Xue X, Chang SF, Fast semantic diffusion for large-scale context-based image and video annotation. IEEE Trans on Image Proc 2012; 21(6):3080–3091. doi:10.1109/TIP.2012.2188038.
48. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al., Imagenet large scale visual recognition challenge. International Journal of Computer Vision 2015; 115(3):211–252. URL: http://dx.doi.org/10.1007/s11263-015-0816-y, doi:10.1007/s11263-015-0816-y.
49. Rendle S, Schmidt-Thieme L, Pairwise interaction tensor factorization for personalized tag recommendation. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, 2010, pp. 81–90.
50. P Wang AFS, Gurrin C, Factorizing time-aware multi-way tensors for enhancing semantic wearable sensing. In: MultiMedia Modeling, MMM 2015, 2015, pp. 571–582.
51. Lee D, Seung H, Learning the parts of objects by nonnegative matrix factorization. Nature 1999; 401:788–791.
52. Shashua A, Hazan T, Non-negative tensor factorization with applications to statistics and computer vision. In: In Proceedings of the International Conference on Machine Learning, ICML, 2005, pp. 792–799.
53. D Xu P Cui WZ, Yang S, Find you from your friends: Graph-based residence location prediction for users in social media. In: ICME 2014, 2014, pp. 1–6.
54. C C Tan Y-G Jiang CWN, Towards textually describing complex video contents with audio-visual concept classifiers. In: ACM Multimedia'11, 2011, pp. 655–658.
55. Merler M, Huang B, Xie L, Hua G, Natsev A, Semantic model vectors for complex video event recognition. IEEE Transactions on Multimedia 2012; 14(1):88–101.
56. Wang P, Sun L, Yang S, Smeaton AF, What are the Limits to Time Series Based Recognition of Semantic Concepts? Cham: Springer International Publishing. ISBN 978-3-319-27674-8, 2016; pp. 277–289, doi:10.1007/978-3-319-27674-8_25. URL: http://dx.doi.org/10.1007/978-3-319-27674-8_25.
57. Doherty AR, Caprani N, O'Conaire C, Kalnikaite V, Gurrin C, O'Connor NE, et al., Passively recognising human activities through lifelogging. Computers in Human Behavior 2011; 27(5):1948–1958.
58. J Guo D Scott FH, Gurrin C, Detecting complex events in user-generated video using concept classifiers. In: CBMI 12, 2012, pp. 1–6.

59. J Liu Q Yu OJSAATADHC, Sawhney H, Video event recognition using concept attributes. In: WACV 2013, 2013, pp. 339–346.
60. S Bhattacharya M Kalayeh RS, Shah M, Recognition of complex events exploiting temporal dynamics between underlying concepts. In: CVPR 2014, 2014.
61. Pirsiavash H RD, Detecting activities of daily living in first-person camera views. In: CVPR 2012, 2012, pp. 2847–2854.
62. Wang P, Sun L, Yang S, Smeaton AF, Gurrin C, Characterizing everyday activities from visual lifelogs based on enhancing concept representation. Computer Vision and Image Understanding 2016; 148:181 – 192. URL: http://www.sciencedirect.com/science/article/pii/S107731421500209X, doi:http://dx.doi.org/10.1016/j.cviu.2015.09.014.
63. Sun C, Nevatia R, Active: Activity concept transitions in video event classification. In: ICCV 2013, 2013, pp. pages 913–920.
64. W Li Q Yu HS, Vasconcelos N, Recognizing activities via bag of words for attribute dynamics. In: CVPR 2013, 2013, pp. 2587–2594.
65. Laptev I, Marszalek M, Schmid C, Rozenfeld B, Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
66. Jaakkola TS, Haussler D, Exploiting generative models in discriminative classifiers. In: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II, Cambridge, MA, USA: MIT Press, 1999, pp. 487–493. ISBN 0-262-11245-0. URL: http://dl.acm.org/citation.cfm?id=340534.340715.
67. Mettes P, Koelma DC, Snoek CG, The imagenet shuffle: Reorganized pre-training for video event detection. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR '16, New York, NY, USA: ACM, 2016, pp. 175–182. ISBN 978-1-4503-4359-6, doi:10.1145/2911996.2912036. URL: http://doi.acm.org/10.1145/2911996.2912036.
68. Whittaker S, Kalnikaitė V, Petrelli D, Sellen A, Villar N, Bergman O, et al., Socio-technical lifelogging: Deriving design principles for a future proof digital past. Human-Computer Interaction (Special Issue on Designing for Personal Memories) 2012; 27(1-2):37.62.
69. Hopfgartner F, Yang Y, Zhou L, Gurrin C, Semantic Models for Adaptive Interactive Systems, chap. User Interaction Templates for the Design of Lifelogging Systems. Springer London, 2013; pp. 187–204.
70. Byrne D, Lee H, Jones GJF, Smeaton AF, Guidelines for the presentation and visualisation of lifelog content. In: iHCI 2008 - Irish HCI 2008, 2008.
71. van den Hoven E, A future-proof past: Designing for remembering experiences. Memory Studies 2014; 7(3):370–384. URL: http://dx.doi.org/10.1177/1750698014530625, doi:10.1177/1750698014530625, http://dx.doi.org/10.1177/1750698014530625.
72. Stix G, Photographic memory: Wearable cam could help patients stave off effects of impaired recall. Scientific American 2011; .
73. Doherty A, Smeaton A, Automatically segmenting lifelog data into events. In: Ninth International Workshop on Image Analysis for Multimedia Interactive Services, IEEE, 2008, pp. 20–23.
74. Zhou L, Caprani N, Gurrin C, O'Connor NE, ShareDay: A novel lifelog management system for group sharing. In: Li S, Saddik A, Wang M, Mei T, Sebe N, Yan S, et al., editors, Advances in Multimedia Modeling, Lecture Notes in Computer Science, vol. 7733, Lecture Notes in Computer Science, vol. 7733, Springer Berlin Heidelberg, 2013; pp. 490–492. ISBN 978-3-642-35727-5, doi:10.1007/978-3-642-35728-2_47. http://dx.doi.org/10.1007/978-3-642-35728-2_47.
75. Naaman M, Harada S, Wang Q, Garcia-Molina H, Paepcke A, Context data in geo-referenced digital photo collections. In: MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA: ACM, 2004, pp. 196–203. ISBN 1-58113-893-8, doi:http://doi.acm.org/10.1145/1027527.1027573.
76. Kitamura K, Yamasaki T, Aizawa K, Food log by analyzing food images. In: Proceedings of the 16th ACM international conference on Multimedia, MM '08, New York, NY, USA: ACM, 2008, pp. 999–1000. ISBN 978-1-60558-303-7, doi:10.1145/1459359.1459548. URL: http://doi.acm.org/10.1145/1459359.1459548.
77. Hughes M, Newman E, Smeaton AF, O'Connor NE, A lifelogging approach to automated market

research. In: Proceedings of the SenseCam Symposium 2012, 2012.
78. Xia L, Ma Y, Fan W, Vtir at the ntcir-12 2016 lifelog semantic access task. In: Proceedings of NTCIR-12, Tokyo, Japan, 2016.
79. Lin HL, Chiang TC, Chen LP, Yang PC, Image searching by events with deep learning for ntcir-12 lifelog. In: Proceedings of NTCIR-12, Tokyo, Japan, 2016.
80. Safadi B, Mulhem P, Qunot G, Chevallet JP, Mrim-lig at ntcir lifelog semantic access task. In: Proceedings of NTCIR-12, Tokyo, Japan, 2016.
81. Scells H, Zuccon G, Kitto K, Qut at the ntcir lifelog semantic access task. In: Proceedings of NTCIR-12, Tokyo, Japan, 2016.
82. de Oliveira Barra G, Ayala AC, Bolaos M, Dimiccoli M, Aghaei M, Carn M, et al., Lemore: A lifelog engine for moments retrieval at the ntcir-lifelog lsat task. In: Proceedings of NTCIR-12, Tokyo, Japan, 2016.
83. Gurrin C, Joho H, Hopfgartner F, Zhou L, Albatal R, Overview of ntcir-12 lifelog task. 2016. The authors acknowledge the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289 and the input of the DCU ethics committee and the risk & compliance officer. We acknowledge financial support by the European Science Foundation via its Research Network Programme ?Evaluating Information Access Systems?., URL: http://eprints.gla.ac.uk/131460/.
84. Caprani N, Doherty AR, Lee H, Smeaton AF, O'Connor NE, Gurrin C, Designing a touch-screen SenseCam browser to support an aging population. In: CHI '10 Extended Abstracts on Human Factors in Computing Systems, CHI EA '10, New York, NY, USA: ACM, 2010, pp. 4291–4296. ISBN 978-1-60558-930-5, doi:10.1145/1753846.1754141. URL: http://doi.acm.org/10.1145/1753846.1754141.
85. Gurrin C, Lee H, Caprani N, Zhang Z, O'Connor N, Carthy D, Browsing large personal multimedia archives in a lean-back environment. In: Advances in Multimedia Modeling, Lecture Notes in Computer Science, vol. Boll, Susanne and Tian, Qi and Zhang, Lei and Zhang, Zili and Chen, Yi-PingPhoebe, Lecture Notes in Computer Science, vol. Boll, Susanne and Tian, Qi and Zhang, Lei and Zhang, Zili and Chen, Yi-PingPhoebe, Springer Berlin Heidelberg, 2010, pp. 98–109.
86. Kelly P, Doherty AR, Smeaton AF, Gurrin C, O'Connor NE, The colour of life: novel visualisations of population lifestyles. In: MM '10 Proceedings of the international conference on Multimedia, ACM, 2010.
87. Redmon J, Divvala S, Girshick R, Farhadi A, undefined, undefined, et al., You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016; 00:779–788. doi:doi.ieeecomputersociety.org/10.1109/CVPR.2016.91.
88. C D, M M, Fergus P, Jones D, Bouhafs F, Exploiting linked data to create rich human digital memories. Computer Communications 2013; 36:1639–1656.
89. Li Y, Guo Y, Wiki-health: From quantified self to self-understanding. Future Generation Computer Systems 2016; 56:333–359.
90. Wang P, Measuring activities of daily living: Digitally, visually and semantically. In: Proceedings of Measuring Behavior 2016: 10th International Conference on Methods and Techniques in Behavioral Research, 2016, pp. 549–556. ISBN 978-1-873769-59-1.
91. Pan SJ, Yang Q, A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 2010; 22(10):13451359.
92. Wang P, Smeaton AF, Semantics-based selection of everyday concepts in visual lifelogging. International Journal of Multimedia Information Retrieval 2012; 1(2):87–101.
93. Wang P, Sun L, Yang S, Smeaton AF, Semantically Smoothed Refinement for Everyday Concept Indexing. Springer International Publishing, 2016; pp. 318–327.
94. Ye G, Li Y, Xu H, Liu D, Chang SF, Eventnet: A large scale structured concept library for complex event detection in video. In: Proceedings of the 23rd ACM International Conference on Multimedia, MM '15, New York, NY, USA: ACM, 2015, pp. 471–480. ISBN 978-1-4503-3459-4, doi:10.1145/2733373.2806221. URL: http://doi.acm.org/10.1145/2733373.2806221.
95. Vinyals O, Toshev A, Bengio S, Erhan D, Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156–3164.

96. Ye T, Moynagh B, Albatal R, Gurrin C, Negative faceblurring: A privacy-by-design approach to visual lifelogging with google glass. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, New York, NY, USA: ACM, 2014, pp. 2036–2038. ISBN 978-1-4503-2598-1, doi:10.1145/2661829.2661841. URL: http://doi.acm.org/10.1145/2661829.2661841.
97. Langheinrich M, Privacy by design - principles of privacy-aware ubiquitous systems. In: Proceedings of the 3rd International Conference on Ubiquitous Computing, UbiComp '01, London, UK, UK: Springer-Verlag, 2001, pp. 273–291. ISBN 3-540-42614-0. URL: http://dl.acm.org/citation.cfm?id=647987.741336.
98. Cavoukian A, Privacy by design: The 7 foundational principles. 2010.
99. P K, SJ M, H B, J K, M O, AR D, et al., An ethical framework for automated, wearable cameras in health behavior research. American Journal of Preventive Medicine 2013; 44(3):314–319.