Using Visual Lifelogs to Automatically Characterise Everyday Activities

Peng Wang, Alan F. Smeaton*

CLARITY: Centre for Sensor Web Technologies and
School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland

Abstract

Visual lifelogging is the term used to describe recording our everyday lives using wearable cameras, for applications which are personal to us and do not involve sharing our recorded data. Current applications of visual lifelogging are built around remembrance or searching for specific events from the past. The purpose of the work reported here is to extend this to allow us to characterise and measure the occurrence of everyday activities of the wearer and in so doing to gain insights into the wearer's everyday behaviour. The methods we use are to capture everyday activities using a wearable camera called SenseCam, and to use an algorithm we have developed which indexes lifelog images by the occurrence of basic semantic concepts. We then use data reduction techniques to automatically generate a profile of the wearer's everyday behaviour and activities. Our algorithm has been evaluated on a large set of concepts investigated from 13 users in a user experiment, and for a group of 16 popular everyday activities we achieve an

*Corresponding author at: CLARITY: Centre for Sensor Web Technologies and School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland. Tel: +353 -1 -7005262

Email address: alan.smeaton@dcu.ie (Alan F. Smeaton)

average F-score of 0.90. Our conclusions are that the the technique we have presented for unobtrusively and ambiently characterising everyday behaviour and activities across individuals is of sufficient accuracy to be usable in a range of applications.

Keywords: Patient monitoring, Wearable camera, SenseCam, Activity classification, Lifelogging

1. Introduction

One of the most significant technological developments in our modern world is the development, and deployment, of various types of sensors throughout our society. As a result we can monitor, in an ambient way, many aspects of our lives and our environment. We are particularly interested in *wearable* sensors which include sensors to directly monitor human behaviour as well as sensors in the mobile devices that we can carry around with us.

Lifelogging is the term 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. This is opposed to having somebody else record what we are doing and using the logged data for some public or shared purpose. For example, an athlete recording his or her daily training and logging distance, time, etc. would count as a form of lifelog whereas a security firm monitoring a train station to detect anti-social behaviour, would not.

One class of personal lifelogging called *visual lifelogging* is based on using wearable cameras, of which there are several examples now available including SenseCam [14], Looxcie®, GoPro®, Vicon Revue and the recently-announced Memento. These record either still images or video and are taken from a first-person view meaning they reflect the viewpoint that the wearer normally sees and usually they also record data from other wearable sensors. Applications for visual lifelogging are manyfold. Sometimes they are job- or task-related, sometimes we lifelog for leisure, and more recently we're seeing lifelogging used for health applications as well as applications for real time lifelogging [34].

In this paper we use visual lifelogging in the task of characterising the activities of the user of a wearable camera, SenseCam. The major contribution of this paper is the proposal of HMM-based models to characterize everyday activities by merging time-varying dynamics of high-level features (concepts). Utilizing concept-concept relationships to map concept vectors to a more compact space by LSA and the extensive experiments on various concept detection performances generated by Monte Carlo simulations are another contribution of this paper, which have not been reported before in semantic analysis of visual lifelogging, to the best of our knowledge. We describe related work in lifelogging followed by an overview of the approaches to managing lifelog data. In section 3 we introduce the background to our experiments including how we define a vocabulary of 85 base semantic concepts. Section 4 presents a Hidden Markov Model approach to deal with the activities to be detected. We then present our experiments using both 'clean' or perfect annotation as well as automatic detection of base semantic concepts, followed by an analysis of two different sampling methods. We conclude the paper with a re-cap on our contribution and directions for future work.

2. Related work

Lifelogging is a very broad topic both in the technologies that can be used, and the applications for lifelogged data. For the most part, lifelogging applications are based around health and wellness, though we have seen applications as diverse as theatre and dance [48]. We describe related work in visual lifelogging where we broadly divide this into applications for memory recall and applications for lifestyle analysis, though new application areas are emerging.

The seminal work in visual lifelogging for memory recall as part of a medical treatment was reported by Berry et al. [4] who used the SenseCam to record personally experienced events so that SenseCam pictures would form a pictorial diary to cue and consolidate autobiographical memories. The SenseCam is a sensor-augmented wearable camera designed to capture a digital record of the wearer's day (as shown in Figure 1) by recording a series of images and other sensor data. By default, images are taken at the rate of about one every 50 seconds while the onboard sensors can help to trigger the capture of pictures when sudden changes are detected in the environment of the wearer. This was used as a form of reminiscence therapy for research into memory and dementia where the visual lifelog acted as a stimulus for recall of short-term events rather than as a memory prosthesis or substitute. Since that initial work many teams have explored the clinical application of visual lifelogging, including Lee and Dey [25] who capture photos, ambient audio, and location information to create a summary of the wearer's recent life for stimulating short-term recall. Another good example of this is the work by Browne et al. [5] 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" [5].

Besides research into the clinical observation of *how* visual lifelogging helps recall, there is also research into *why* this seems to work. St. Jacques

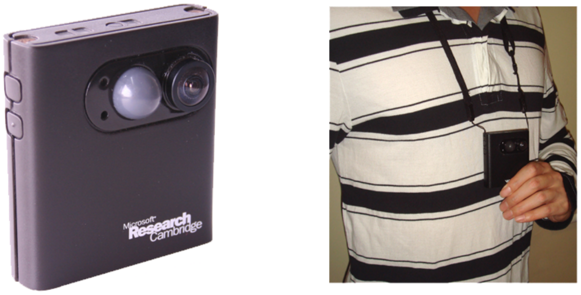


Figure 1: The Microsoft SenseCam (right as worn by a user).

et al. [42] has experimented with using fMRI scans to observe which parts of the brain are most active when subjects are benefitting from the stimulating effects of their own visual lifelogs and while this work and the similar work of others is still ongoing, it is observed that viewing the visual lifelog of one's own recent past stimulates and opens the pathways for autobiographical recall which may otherwise have been blocked or closed as a result of memory impairments caused by acquired brain injury or various forms of early-stage dementia. How to keep these pathways open or understanding how this happens is not yet known or understood.

Visual lifelogging has also been used to analyse lifestyles and behaviours in individuals and in groups. This is typically done by developing automatic classifiers or software detectors for individual base concepts or even for objects, and then applying these to visual lifelogs to infer different lifestyle traits or characteristics Doherty et al. [9]. The accuracy of these automatic detectors will vary because it is a difficult problem to overcome anyway and, for example, determining whether a wearer was actually outdoors or just looking out a window from indoors, or recognising the difference between a wearer eating or just preparing food, is very challenging.

In addition to characterising overall activities, visual lifelogging can also be used to analyse and characterise specific activities. For example, recent results from a pilot study have examined both the sedentary and the travel behaviour of a population of users and early results have shown considerable potential in the field of travel research where the duration of journeys to/from work or school, as a specific targeted activity, can be accurately estimated just from SenseCam images [20]. Other approaches to analysis of human behaviour in the home such as by [37] have required expensive investment in infrastructure in the home whereas by using wearable sensors, our work is more scalable.

What all this related work has in common with ours is that visual lifelogs are large and often unstructured collections of multimedia information and there is a real problem across all applications of how to access this information in a structured way. To make progress on this, naturally we turn to any relevant theories in the area to help and guide us here and if there is guidance it should come from the area of information science. Unfortunately, information science has not (yet) addressed the challenges of how to manage information access to lifelogs, possibly because the area is still so young so we default and use whatever guidance we can find from information access to less complete personal media, such as our personal photos, for example. In managing personal media we find that some metadata like date, time and perhaps even location, may help but what we really want is access to visual lifelogs based on their content rather than their metadata. In the next section we look more closely at how access to lifelog data can be managed.

3. Managing Lifelog Data

The application of lifelogging, especially visual lifelogging, to analysis of the activities of the wearer creates challenging problems for retrieval due to the large volume of lifelog data and the fact that much of that data is repetitive with repeated images of the same or nearly the same thing. Recording every activity of a wearer's life will generate a large amount of data for a typical day, not to say for a longer term, for example, a month or even a year. Detecting events or activities which are distinct from such a mass of homogeneous lifelog data without efficient indexing and retrieval tools, poses a real problem.

Our approach to managing lifelog data is to index visual lifelogs on simple or *"base"* semantic concepts automatically detected from within the lifelog images. The challenges here include defining which base concepts we should try to detect as well as determining which methods we should use in order to achieve high accuracy in a computationally efficient way.

3.1. Concept-based Lifelog Retrieval

Concept-based lifelog retrieval has not received much attention from the lifelogging community because it is such a new and emergent area. Using automatically-detected simple or base semantic concepts for multimedia retrieval has attracted interest in other domains however, including broadcast TV news video, movies and surveillance video [40, 39]. In principle this is very attractive – a concept is either present, or absent, from an image or video clip which makes retrieval a straightforward binary search – but in practice there is always a degree of (un)certainty associated with the detection, which we can regard as a probability of the concept being present. To further complicate things, the actual detection process is not perfect and so it will have an accuracy or effectiveness level, on top of a probability of occurrence.

Toharia et al. [45] have artificially varied the level of detection accuracy to determine the "tipping point" for detection accuracy above which semantic concepts become useful aids in multimedia retrieval. With continuing progress in the development of techniques for automatic concept detection in various domains we have now reached this point of being able to achieve satisfactory results. For example, Aly et al. [1] worked in the domain of TV news video where semantic concepts can be detected directly from the video and they developed and tested a probabilistic model which accounts for the uncertain presence of such concepts. This was then evaluated in the application of retrieving news stories from TV news broadcasts. In [13], the semantic model vector (the output of concept detectors) has already been shown to be the best-performing single feature for IBM's multimedia event detection task in the annual TRECVid video retrieval benchmarking evaluation [38]. This whole approach has now moved to the point where we could develop techniques to then *fuse* the automatically detected base concepts into higher level semantics and achieve levels of indexing for which the usual methods are not capable.

According to research results in the neuroscience area, we tend to remember our past experiences when structured in the form of events [51]. This poses a requirement for lifelogging tools to provide high-level concept detection facilities to categorize lifelog data, including images, into events for organization or for use as a re-experience. Such real events as sitting on a bus, walking to a restaurant, eating a meal, watching TV, etc. consist of many, usually hundreds, of individual images taken from a wearable camera such as SenseCam. In many cases, where the wearer is moving around, a large number of images which are quite dissimilar in appearance, may be generated. The variety of SenseCam images in lifelogging introduces challenges for event detection when compared to traditional TV news broadcasting video, for example. The image capture rate which is not continuous, also makes dynamic spatial-temporal features like HOG (Histograms of Oriented Gradients) and HOF (Histograms of Optical Flow) descriptors, inapplicable because frame-to-frame differences can be so large whereas when we classify moving video into events, such approaches are well-matched and quite well developed [19, 16, 13]. In lifelogging, when concept detection is usually carried out it is at the level of individual images rather than at the event level, which is where it is required. The variety of concepts within one event makes event detection and event categorization difficult and this is the problem we tackle here. We now describe how to apply concept-based indexing of visual lifelogs using a new approach of classifying everyday activities.

3.2. Analysing Activities from Visual Lifelogs

In our research, the problem of detecting events and base concepts from lifelog images taken by a SenseCam wearable camera, is simplified into a classification problem, that is, we find the most likely concept(s) for an event from a lexicon set with regard to the event input.

Suppose we are given an annotated training set $\{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$ consisting of N independent examples. Each of the examples $x^{(i)}$ represents the i-th event in the corpus. The corresponding annotation $y^{(i)} \in [1, |T|]$

is one of the concepts in the lexicon T. The task for event-based concept detection can be described as: given the training set, to learn a function $h : \mathcal{X} \mapsto \mathcal{Y}$ so that $h(x)$ is a predictor with an unlabeled event input x for the corresponding value of y.

Progressing through the concept detection procedure, each image is assigned labels indicating if specific concepts are likely to exist in the image or not. If we have the universe of concept detector set C, event $x^{(i)}$ is represented by successive images $I^{(i)} = \{Im_1^{(i)}, Im_2^{(i)} ... Im_m^{(i)}\}$. The concept detection result for image $Im_j^{(i)}$ can be represented as an n-dimensional concept vector, as $C_j^{(i)} = (c_{j1}^{(i)}, c_{j2}^{(i)} ... c_{jn}^{(i)})^T$, where n is equal to the cardinality of C and $c_{jk}^{(i)} = 1$ if concept k is detected in the image, otherwise $c_{jk}^{(i)} = 0$.

While the SenseCam wearer is performing an activity which requires him/her to be moving around, the first-person view may change over time, though not, for example, if he/she is watching TV or sitting in an office looking at a computer. We need to map time-varying concept patterns from individual images into different activities for events. To classify an event consisting of a series of images in temporal order is very similar to recognizing a phoneme in an an acoustic stream, to some extent. The event is analogous to a phoneme in the stream and every image within this event is analogous to an acoustic frame. The task of temporal activity classification is thus suitable to be addressed by a classical Hidden Markov Model (HMM), which has proved to be efficient in speech recognition [35]. In addition to the systematic theory of HMM and many successful applications in computer vision, its powerful framework for temporal modeling of time-varying features makes it very flexible when applied to this kind of problem like event recognition, especially when cardinality (i.e. the length of the stream) differs across different event segments and is unforseen. HMM can be adaptive in this situation and hence avoid extra processing like time warping as used in [7], or vector quantization as used in [16], which is necessary for other learning algorithms demanding feature vectors of fixed dimension (such as SVM and KNN). Due to its intrinsic advantages, HMM has been widely adopted in multimedia retrieval area by much research on classification or recognition tasks such as [50], [10], [12], [27], [13] and [19], just to name a few, though different features or forms are applied. We now elaborate the construction of HMMs for the solution to our problem.

3.3. Selecting Base Concepts for Activity Analysis

Before automatically categorizing everyday activities, we first need to explore the activities to be analyzed in building and testing our algorithm. Theories from psychology can guide our selection of human activities and then the subsequent set of related concepts which can help us to construct a meaningful semantic system for our classification task. Maslow's hierarchy of needs [28, 31] is one such theory of human motivation, organizing different levels of needs in the shape of a pyramid, with the largest and most fundamental levels of need at the bottom, and the need for self-actualization at the top. This theory analyzed the relationships between various needs and their impact on human motivation. Another source of guidance for choosing target activities is taken from occupational therapy, in which the occupation and its influence on health and well-being are studied as one important topic. The relationship between our engagement in everyday activities and well-being for individuals has been shown and evidence has demonstrated the existence of this relationship within various age groups [24, 29]. As improving human health and well-being is an important goal of lifelogging research, we borrowed more outcomes from occupational therapy research and applied them into the selection of our target activities and concepts. Applying Maslow's motivation theory might suffer from a deficiency in finding activities with high enough generality across groups of people with different backgrounds since the hierarchy may vary with culture [11, 21] or circumstance [43, 44].

Investigations and surveys in the area of occupational therapy have shown that most of our time is spent on activities such as sleeping and resting (34%), domestic activities (13%), TV/radio/music/computers (11%), eating and drinking (9%), which collectively count for nearly 70% of the time in a typical day [6]. For example, according to the UK Government, Office for National Statistics (ONS) [46], [47] and Chilvers et al. [6], some activities like "sleeping", "housework", "watching TV", "employment/study", "eating/drinking", etc., are the most prevalent activities on which most of the time in our daily lives is spent. Some of these activities like "sleeping", "eating/drinking", "personal care", "travel", etc., are engaged across all age groups and thus achieve very high participation agreement among all people investigated in those surveys.

Though there exist numerous activities in our everyday lives, as demonstrated above in occupational therapy research, recent work utilizing wearable computing devices to improve human health tries to concentrate on those activities which are more frequent or prevalent in our lives. For example, eating is the focus of diet monitoring [36, 17] while instrumental daily activities such as "making coffee", "cooking", etc., are targets in analysis for the diagnosis of dementia [30]. In [18], a similar set of 16 everyday activities is explored to rate their level of enjoyment when people experience these activities. The impact of everyday activities on our feelings of enjoyment also affect our health, which makes these activities important in an analysis of well-being and lifelogging which is our focus. An investigation into the automatic detection of these activities would be of great value for medical applications like obesity analysis, and chronic disease diagnosis, to name just two. After carefully considering the activities we discussed above, we selected 23 activities as targets for our further experiment and analysis which are listed in Table 1, which have shown their effectiveness in validating manual and automatic concept selection methods in [49]. These activities were chosen based on the following criteria:

Table 1: Target activities for our lifelogging work

1 Eating	2 Drinking	3 Cooking	4 Clean/Tidy/Wash
5 Wash clothes	6 Using computer	7 Watch TV	8 Children care
9 Food shopping	10 General Shopping	11 Bar/Pub	12 Using phone
13 Reading	14 Cycling	15 Pet care	16 Go to cinema
17 Driving	18 Taking bus	19 Walking	20 Meeting
21 Presentation (give)	22 Presentation (listen)	23 Talking	

- Time dominance: As described above, a small number of activities occupy a large amount of our time. The analysis of these activities can maximize the analysis of the relationship between time spent and human health. The selected activities should collectively cover most of the time spent in a day.

- Generality: Even though the time spent on activities varies from age group to age group, there are some activities that are engaged in by multiple age groups. The selection of activities with high group agreement will increase the generality of our activity analysis in lifelogging so the output can be suited to a wider range of age groups.

- High frequency: This criteria helps to select activities which have enough sample data in our lifelogging records. High sample frequency can improve the detection and other processing qualities, such as classification and interpretation.

Note that detecting base concepts is the first step when performing activity classification. Among the efforts in building semantic lexicons for concept-based retrieval, the Large-Scale Concept Ontology for Multimedia (LSCOM) is the most comprehensive taxonomy developed for standardizing multimedia semantics in the broadcast TV news domain [32]. The LSCOM consortium tried to bring together experts from multiple communities including multimedia, ontology engineering and others with domain expertise. Multiple criteria were also considered including utility, coverage, etc. in the selection procedure. In order to systematically examine automated concept detection methods, Snoek *et al.* manually extended both the number of concepts and the number of annotations by browsing the training video shots for TV news broadcast programmes [41]. The manual annotation process finally yielded a pool of ground truth for a lexicon of 101 semantic concepts. Since there is no theoretical way to precisely define the lexicon for a given domain, manual concept selection is an important approach for the evaluation of automatic concept detection and thus concept-based retrieval methods. This can also be shown by the light scale multimedia concept ontology which is also employed in TRECVid 2005 [33]. In [15], manual selection of concepts for query tasks is presented as one benchmark, together with the comparison with the other benchmarks generated from an extensively tagged collection based on mutual information [26]. A human-generated ontology of everyday concepts is employed in [49] and compared with a density-based concept selection approach through semantic reasoning.

To utilize concepts to reflect domain semantics appropriately, a user experiment was carried out to identify topic-related concepts with regard to activities listed in Table 1, as presented in [49]. Although individuals may have different contexts and personal characteristics, a common understanding of concepts that is already socially constructed and allows people to communicate according to Lakoff [22] and Huurnink et al. [15], also makes it possible for users to define suitable base concepts which are relevant to activities.

A total of 13 respondents took part in our user experiment, chosen from among researchers within our own research group, most of whom are working in computer science and some of them also log their own everyday lives with the SenseCam. This means the group are sympathetic to and familiar with the idea of indexing visual content by semantic concepts. In the experiment, target activities were first described to the respondents to make them familiar with the activity. Participants were then shown SenseCam images for selected activity examples and surveyed by questionnaire about their common interpretation of the SenseCam activity images as well as of the concepts occurring regularly in those SenseCam images. The aim of the user experiment was to determine candidate semantic concepts which have high correlation with human activity semantics. After several iterations and refinements we selected 85 base concepts which have the highest agreement among respondents, i.e. more than half of respondents think each of them are relevant to the underlying activity. These 85 concepts described in Table 2 as a universal set organized into general categories of objects, scene/setting/site, people and events, were then employed as the base concepts for the rest of the work in this paper. Note that these 85 concepts are borrowed as the small concept set from [49] and more details about the user experiment and the use of this concept ontology can be found in [49].

4. Experimental Setup and Variables

In this paper, the methodology we proposed for the investigation of everyday activity characterization can be demonstrated by the algorithm pipeline shown in Figure 2. The algorithm consists of four main components which

Table 2: Set of 85 Experimental Concepts

Objects	*plate, cup, cutlery, bowl, glass, bottle, milk, drink, fridge, microwave, cooker, water, cloth, clothes, glove, soap, hanger, screen, keyboard, monitor, TV, remote control, basket, trolley, plastic bag, mobile phone, phone screen, book, newspaper, notebook, paper, handle bar, steering wheel, car, bus, bicycle, pet, road sign, traffic light, cat, yellow pole, chair, laptop, projector, pram/buggy*
Scene/ Settings/ Site	*indoor, outdoor, office, kitchen, table, sink, basin, toys, shelf, cashier, door, building, fruit, vegetable, deli, food, road, path, cycle lane, sky, tree, dark, window, inside bus, shop, inside car, projection*
People	*face, people, group, child, hand, finger*
Event	*hand washing, hanging clothes, hand gesture, finger touch, page turning, presentation, taking notes*

are concept identification from raw SenseCam images, vocabulary construction for visual semantics, the modeling of time-varying patterns by HMM and activity classification through trained HMM models. The vocabulary construction module can further be boiled down to LSA (Latent Semantic Analysis) and vector quantization, which we will discuss in more detail later in this section. The activity modeling using HMM and the training of parameters will also be elaborated in this section and the simulation of concept detection results will be discussed in Section 5 when varying concept detection accuracy, aiming to drill down the analysis of algorithm performance.

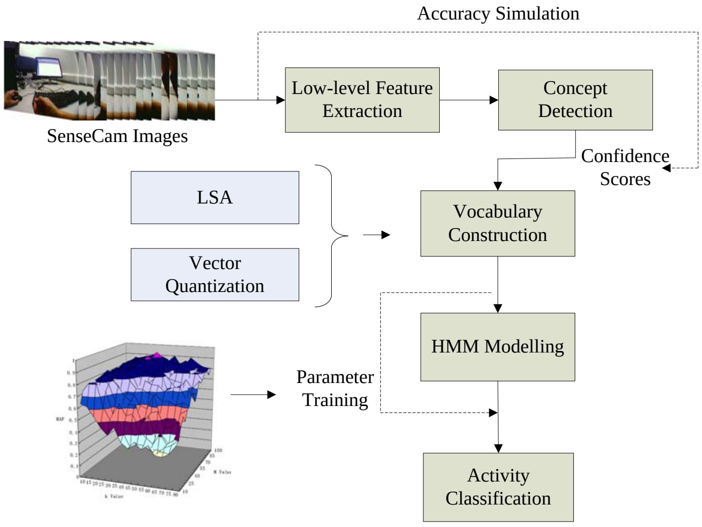


Figure 2: The algorithm pipeline.

4.1. Vocabulary Construction for SenseCam Images

Concept detection provides us with a way to determine the appearance of concepts in images which can be used as high-level semantic features for later concept-based retrieval or even further statistical classification. Concepts can play different roles in representing event semantics, and some will interact with each other through their ontological relationships. This means that if we plot the concepts in a vector space, the dimensions in a concept vector $C_j^{(i)}$ are not independent because of their relationships to each other. Ignoring concept-concept relationships would likely degrade the performance of any subsequent classification of activities.

We address the underlying semantic structure using Latent Semantic Analysis (LSA) [8]. In the traditional Vector Space Model, LSA can represent index terms and documents by vectors and can analyze document relationships in terms of the angle between two vectors, each representing a document. The advantage of LSA is that the terms and documents are projected to a concept space and retrieval performance is improved by eliminating the “noise” in the original space. In LSA, the similarity of meaning between terms is determined by a set of mutual constraints provided by term contexts in which a given term does and does not appear [23]. The application of LSA in our research can be described as the following:

Assume that we have n concept detectors and a corpus consisting of m SenseCam images. We can construct an $n \times m$ concept-image matrix:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} \tag{1}$$

where each element $x_{ij} = 1$ if concept c_i appears in image I_j, otherwise $x_{ij} = 0$. In the matrix $\mathbf{X}$, each row represents a unique concept and each column stands for an image. LSA is carried out by applying Singular Value Decomposition (SVD) to the matrix. The concept-image matrix is decomposed into the product of three matrices as shown:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are left and right singular vectors respectively, while $\mathbf{\Sigma}$ is the diagonal singular matrix of scaling values. Both $\mathbf{U}$ and $\mathbf{V}$ have orthogonal columns and describe the original row entities (concepts) and column entities (images) separately. Through SVD, the matrix X can be reconstructed approximately with fewer dimensions $k < n$ using the least squares manner. This can be done by choosing the first k largest singular values in $\mathbf{\Sigma}$ and the corresponding orthogonal columns in $\mathbf{U}$ and $\mathbf{V}$. This yields the approximation as:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{U_k}\mathbf{\Sigma_k}\mathbf{V_k^T} \tag{3}$$

The reduced matrix not only retains the semantic relationship between original base concepts and images, but also removes "noise" induced by base concepts which are similar to each other. Since $\mathbf{U_k}$ is an orthogonal matrix, it is not hard to calculate the projection of any sample vector C_j in the new concept space as:

$$\hat{C_j} = \mathbf{\Sigma_k^{-1}}\mathbf{U_k^T}C_j \tag{4}$$

After the concept vectors are mapped to the new concept space, vector quantization is employed to represent similar vectors with the same index. This is performed by dividing the large set of vectors into groups having a number of points similar to each other. In this way, the sample vectors characterizing concept occurrences are modeled only by a group of discrete states which is referred to as the *vocabulary*. Vector quantization is done by clustering sample sets in an n-dimensional space, to M clusters, where n is the number of space bases (k after LSA), while M is the vocabulary size.

For vector quantization, we applied a k-means clustering algorithm to categorize the samples in the k-dimensional space. To avoid local optimization of quantization error, we carried out 10 iterations of k-means clustering with different randomly initialized cluster centers. The clustering result with minimum square error is selected as the final vocabulary. One example of vocabulary construction is shown in Figure 3, in which sample points are projected in a $2 - d$ concept space and clustered for a vocabulary of size 5.

4.2. HMM Model Structure

In our activity detection, each segmented lifelog event is treated as an instance of an underlying activity type, constructed from a series of SenseCam images. A Hidden Markov Model [35] is a very efficient machine learning tool to model time-varying patterns. In our activity classification, the HMM treats the event instances as mutually independent sets of concepts generated by a latent state in a time series. The model structure as shown in Figure

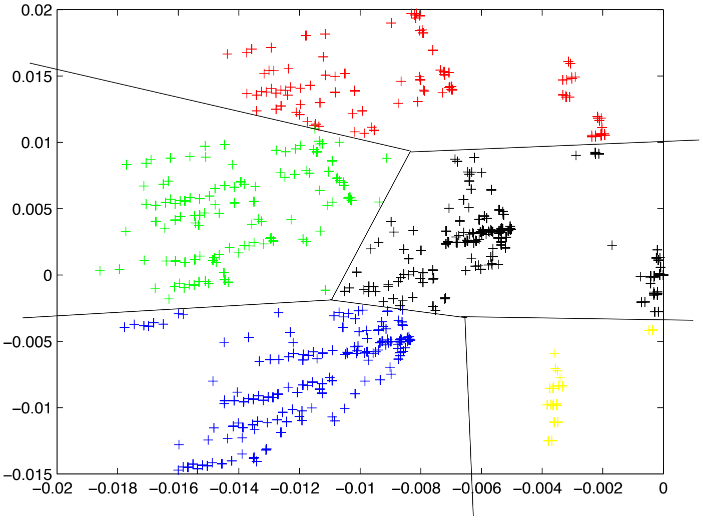


Figure 3: Vocabulary construction example in $2D$ space.

4 is used in modeling the temporal pattern of dynamic concept appearances in an activity.

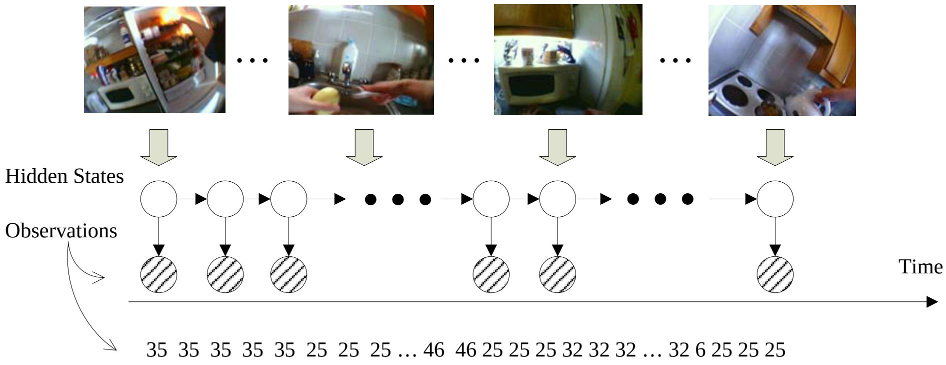


Figure 4: HMM structure for activity modeling.

In Figure 4, one ‘Cooking’ event is demonstrated by the change of states and observation sequences, through the time line. The fully connected state transition model is shown in Figure 5:

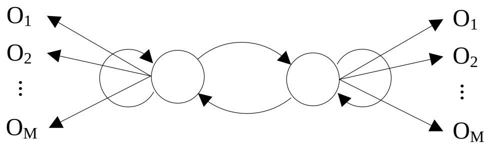


Figure 5: Two states transition model.

4.3. Parameter Training

The choice of k and M, which determine the amount of dimension reduction in the concept space and vocabulary size will affect the performance of our algorithm. The choice of k should be large enough to reflect the real structure in a new concept space but be small enough to avoid sampling errors or unimportant details introduced in the original matrix. It is a similar case in selecting an appropriate value for M for which the representation of observation and modeling complexity should also be balanced. Finding choices of k and the cluster number M in a theoretical way is beyond the scope of our work and is an open issue in the information retrieval and machine learning communities. In our work, we regard k and M as two parameters and test the best combination with the criterion of maximising retrieval performance, namely mean average precision (MAP).

We trained an HMM model for each activity class, that is, for each activity type described earlier we trained the model with multiple observation sequences to find the optimal parameters. This is done using the Baum-Welch algorithm which optimally estimates the probability of the HMM model by iteratively re-estimating model parameters. In our experiment we cross-validated the HMM models on training data with *leave-one-out* cross validation. After a number of iterations, the best-initialized HMM parameters are selected and the HMM model is trained on all training data sets for the activity type. The models for different activity types are then evaluated on the final test data to assess retrieval performance. The detailed model training and parameter searching will be presented in the next section, Section 5.

5. Experimental Results

5.1. Evaluation Data Set

In the experiment on evaluating activity classification, we carried out an assessment of our algorithm on data sets using both clean (correct) concept annotation and on concept annotation with errors. The data sets we used are event samples of the 23 activity types from Section 3.3 collected by 4 people with different demographics (older people vs. younger researchers), one older participant who is less functional in terms of capacity for household and community activities from an occupational therapist's standpoint. The choice of these four people can help to test if our algorithm is applicable from among a group of different people.

Privacy is the first issue to be considered during the preparation for experiment data since the SenseCam can record quite detailed activities uncer certain constraints and conditions. Ethical approval was obtained from Research Ethics Committee in our university for the use of participants' SenseCam images. Another reason why we chose these four people's data in the experiment is that all of them have been wearing the SenseCam consecutively for more than 7 days. This guarantees a variety of activities (for example, 'Eating' at home or outside, 'Walking' in the street or countryside, etc.) needed in our experiment and one week's recording can also better reflect the life pattern of the participants and semantic dynamics of their activities. Note that the SenseCam images of these four participants have very good image quality as the observability of visual concepts is another criterion for us to choose the experiment data. Due to the limited number of positive samples of each activity type, we use 50% of each sample for training and 50% for testing. Event types with more than 5 positive samples are selected, giving the 16 event types shown in Table 3 with sample number and numbers of images contained.

5.2. Evaluation on Clean Concept Annotations

The clean concept annotation means the concept annotations on each image are error-free. This is achieved by manually annotating the 85 base concepts we proposed in Section 3.3 for the data sets. For annotation purposes,

Table 3: Experimental data set for activity classification

Type	Eating	Drinking	Cooking	Clean/Tidy/Wash
# Samples	28	15	9	21
# Images	1,484	188	619	411
Type	Watch TV	Child care	Food shopping	General shopping
# Samples	11	19	13	7
#Images	285	846	633	359
Type	Reading	Driving	Use phone	Taking bus
# Samples	22	20	12	9
# Images	835	1,047	393	526
Type	Walking	Presentation (listen to)	Use computer	Talking
# Samples	19	11	17	17
# Images	672	644	851	704

a concept annotation software tool was developed to inspect the SenseCam images and judge if each concept exists or not. The temporal relationship is kept during annotation by providing a series of SenseCam images within the same event. This helps to improve annotation speed for the user by selecting positive image examples and the unselected samples will be annotated as negative examples. Thus a group of images can be annotated in one click and the whole event can be annotated for one concept in just a few clicks. The performance of activity classification on this clean annotation is now described.

As described in Section 4, the selection of parameters k and M will affect the performance of our algorithm. In our experiment, we evaluated the final retrieval performance with different settings of these parameters. The search graph of parameters k and M in order to tune MAP is shown in Figure 6, for which a 3-state HMM model is used.

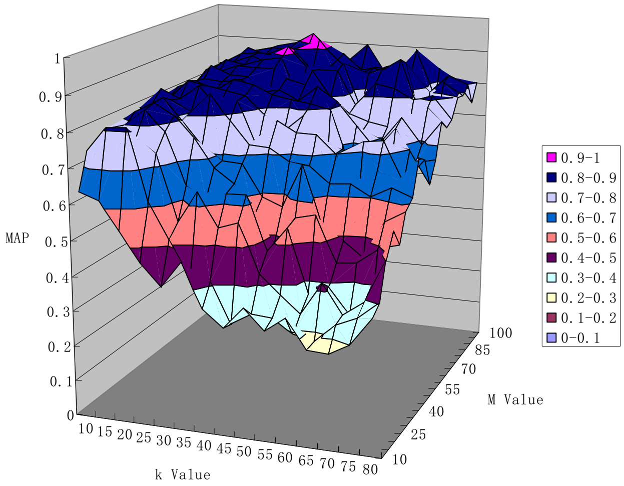


Figure 6: Search graph for MAP optimisation (3 states).

The search graph is built by varying k and M in the ranges [10..80] and [10..100] respectively. The best performances ($MAP \geq 0.9$) appear in the range [30..50] and [80..100] for k and M. When the value of k is increased, the value of M also needs to increase to achieve better performance. The worst case happens when selecting a large k value and small M value, when more 'noise' is introduced from the concept space and the vocabulary clusters can not adapt to this noise. The situation is better when k is low, say, $k = 20$, for which most choices of M have MAP above 0.8. Meanwhile, large M values can also complement the choice of k, when M is large enough ($M \geq 90$), most MAP remain at a satisfactory level, even though the best cases are in the range $k \in [30..50]$. A similar pattern can be seen when choosing different state numbers.

In our training experiment, we trained and tested different settings of model parameters including the dimensions of concept space and the vocabulary size. After testing different combinations, we selected a concept space dimension of 35 and vocabulary size of 80 for further investigation. Different numbers of hidden states were tried in 5 runs and the overall performance (average MAP) is considered in choosing the state number. In our experiment, 2 states achieves best overall performance which is then used to train HMM models for each type of activity. Because each HMM model can return the likelihood of an observation sequence, we perform activity classification by selecting the class of activity with highest likelihood for the input observation. The performance is then evaluated by precision, recall and F-Score as shown in Table 4.

5.3. Concept Detection with Errors

In order to assess the performance of our activity detection algorithm on *automatically* detected rather than manually annotated base concepts which will have some errors in their detection, we manually controlled the simulated concept detection accuracy, based on the groundtruth annotation. The simulation procedure is borrowed from Aly et al. [3], in which they use Monte Carlo simulations to generate various accuracy performances for concept detection.

The notion of this simulation is based on the approximation of confidence

Table 4: Event detection results

Event type	Precision	Recall	F-Score
Child care	0.68	1.00	0.81
Clean/Tidy/Wash	0.86	0.86	0.86
Cooking	0.80	0.89	0.84
Drinking	0.75	0.80	0.77
Driving	1.00	1.00	1.00
Eating	0.95	0.75	0.84
Food shopping	1.00	1.00	1.00
General shopping	0.86	0.86	0.86
Presentation (listen)	1.00	1.00	1.00
Reading	1.00	0.95	0.98
Taking bus	1.00	0.89	0.95
Talking	0.85	0.65	0.73
Use computer	1.00	1.00	1.00
Use phone	0.92	1.00	0.96
Walking	0.86	0.95	0.90
Watch TV	1.00	0.82	0.90

score outputs from concept detectors as a probabilistic model of two Gaussians. In other words, both the densities for the positive and negative classes of a concept are simulated as Gaussian distributions. The concept detector performance is then controlled by modifying the models' parameters [3]. The method also assumes that the confidence scores of different detectors for a single object such as an image are independent from each other. All concepts are assumed to share the same mean μ_1 and standard deviation σ_1 for the positive class while the mean μ_0 and the standard deviation σ_0 are for the negative class. The performance of concept detection is affected by the intersection of the areas under the two probability density curves whose shapes can be controlled by changing the means or the standard deviations of the two classes for a single concept detector.

Our implementation of the simulation involves the following processes. First, we simulate the confidence observations of concept detector as $N(\mu_0, \sigma_0)$ and $N(\mu_1, \sigma_1)$ for the negative and positive classes respectively. The prior probability $P(C)$ for a concept C can also be obtained from the annotated collection. Then the sigmoid posterior probability function with the form of Equation 5 is fit for the generation of a specified number of S training examples.

$$P(C|o) = \frac{1}{1 + exp(Ao + B)} \tag{5}$$

After parameters A and B are decided, the posterior probability of the concept is returned using the sigmoid function for each shot with a random confidence score o drawn from the corresponding normal distribution. A more detailed description of the simulation approach can be found in [3] and [2].

In setting up the "concept detectors with errors" in our experiment, we modified the concept detection performance with the simulation based on the groundtruth annotation described in Table 3, for which each image is annotated with the existence of all concepts. During the simulation procedure, we fixed the two standard deviations and the mean of the negative class. The mean of the positive class was changed to the range of [0.5 .. 10.0] to adjust the intersection area within the two normal curves, thus changing the detection performance. For each setting of parameters, we executed 20 repeated runs and the averaged concept detection MAP and averaged activity detection MAP were both calculated.

5.4. Evaluation on Concept Detection with Errors

The evaluation on erroneous concept detection was carried out by training and testing the activity detection algorithm described in Section 4, on the simulated concept detections with variable detection accuracy. We increased the mean of the positive class μ_1 for each concept in our lexicon from 0.5 to 10.0 with step 0.5. For each value of μ_1, we executed 20 simulation runs, and for each run the concept detection MAP was calculated.

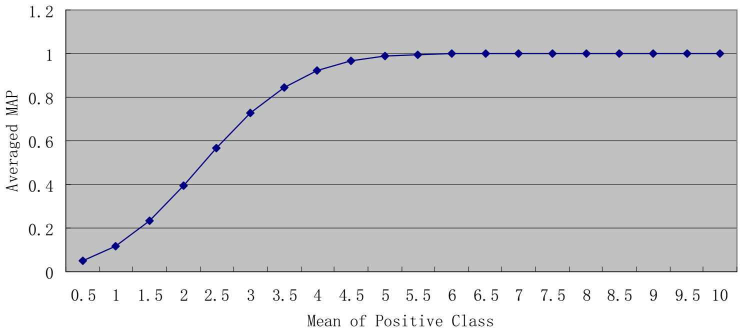


Figure 7: Averaged concept MAP with different positive class means.

In Figure 7, the concept MAP for all 20 runs are averaged and plotted with the increase of positive class mean μ_1. The x-axis shows the changes of μ_1 with the setting of the other parameters as $\sigma_0 = 1.0$, $\sigma_1 = 1.0$ and $\mu_0 = 0.0$. The y-axis depicts the value of averaged concept MAP for each μ_1. From

Figure 7 we can see that increasing μ_1 achieves better concept detection performance. When μ_1 reaches the value 5.5, the concept detectors almost have the same performance as the ground truth and can be regarded as perfect.

For each run, the simulated concept annotations were analyzed by LSA first and projected to a new concept space with a lower dimension of $k = 35$. Vector quantization was then carried out in the new space by k-mean clustering, representing every SenseCam image with one observation from the vocabulary constructed. After vector quantization, the SenseCam image which was formerly represented with a 85-dimensional vector, was indexed with only the number of the cluster. In this step, we still choose $M = 80$ and achieve 80 clusters in the new concept space. The dynamic pattern of observations was modeled by the HMM model whose parameters were trained in the same process as described in Section 4. The testing was performed on the data set described in Section 5.1.

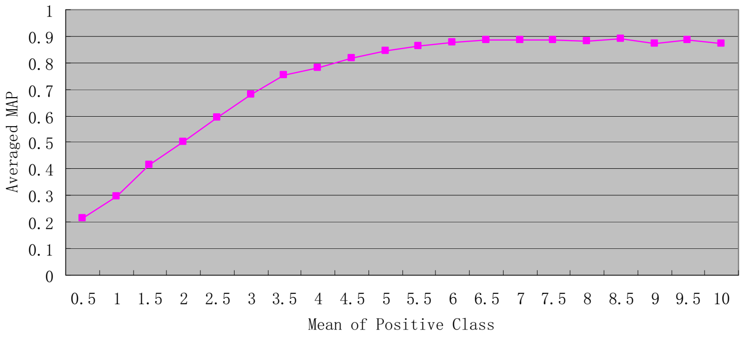


Figure 8: Averaged activity MAP with different positive class means.

Figure 8 depicts the changes in averaged activity detection MAP with respect to the μ_1 values, using the half-and-half sampling method for training and testing data. The x-axis has the same meaning as it has in Figure 7 while the y-axis is the averaged MAP of activity detection over 20 runs. As expected, the activity detection performance increases with improving concept detection performance. Note that the activity detection performance does not drop significantly when the concept MAP is low. The smooth change of activity detection MAP shows that our algorithm is robust and tolerant to the errors introduced in automatic concept detection.

6. Discussion

6.1. Varying the Sampling Method

As described in Section 5.1, each event sample is divided into two halves, of which the first half is used as training data and the other is used as testing data. To evaluate the effect of this sampling method for training and test data, we also carried out the same experiment on another sampling method, odd-and-even sampling, to distinguish from half-and-half sampling. That is, in each event sample, we used the odd numbered images as training data while the images with even number are used as testing data. The performance comparison of the two sampling methods on the clean data set is shown in Figure 9.

For evaluation purposes, the training and testing are carried out for 10 runs with each of the two sampling approaches. During the procedure, we

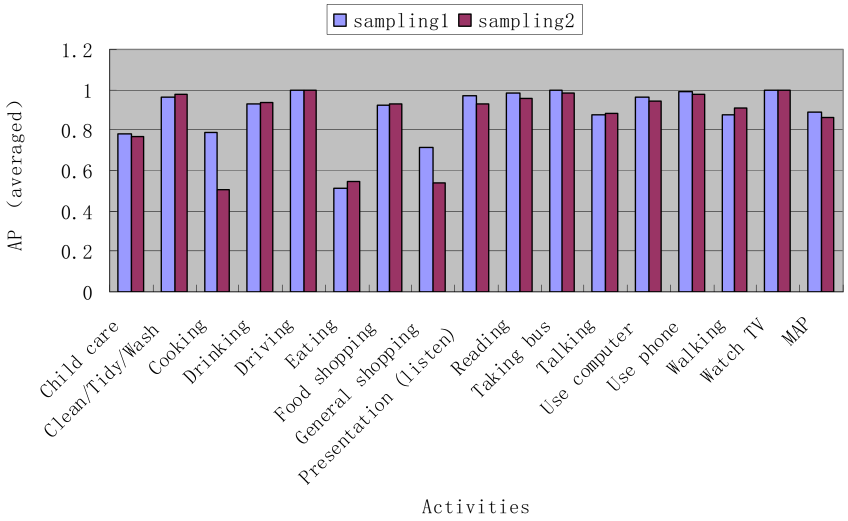


Figure 9: Comparison of two sampling methods (clean data).

used the same parameter settings as above, $k = 35$, $M = 80$, and 2 hidden states. The activity detection AP is calculated for each activity and then averaged on these 10 runs. The two sampling approaches are compared on the activity basis out of the 16 activities investigated. In Figure 9, the averaged AP for each activity and averaged MAP are shown. The half-and-half sampling and odd-and-even sampling are represented as $sampling1$ and $sampling2$ respectively in the figure. From Figure 9, there is no obvious difference between two sampling methods for most activities, compared using AP. Only two activities show obvious performance differences, which are 'Cooking' and 'General shopping'. The drop in performance for odd-and-even sampling shows that this sampling method can disrupt the intrinsic observation transition, especially for activities in which the observation of concepts during the event changes frequently like ‘Cooking’. For those activities in which base concepts appear with greater stability during the event like ‘Driving’, ‘Taking bus’, ‘Watch TV’, etc., the performances of two sampling methods are almost the same. The overall performance also dropped using odd-and-even sampling as reflected by averaged MAP which is 0.89 for $sampling1$ and 0.86 for $sampling2$. The performance difference shows that base concept observation patterns can be changed by the odd-and-even sampling method. On the other hand, this also reflects that our algorithm can capture the pattern of concept dynamics and apply these patterns in activity classification for better performance. The evaluation of two sampling methods on erroneous concept detection is now described.

Similar to using clean concept annotation data, we also compared the two sampling methods on simulated concept detection. This is performed by changing the mean of positive class μ_1 for each concept and 20 runs are carried out for each value of μ_1. For each simulation run, the evaluation procedure involved training and testing steps which are the same as using clean data and we use exactly the same parameter settings. Activity detection MAP is calculated in each run and then averaged on all 20 runs to obtain the overall performance on one simulation configuration. The performances of two sampling methods are shown in Figure 10.

In Figure 10, the x-axis shows the configurations of μ_1, varying from 0.5 to 10.0. The settings for the parameters are the same as in Figure 7, that is $\sigma_0 = 1.0$, $\sigma_1 = 1.0$ and $\mu_0 = 0.0$. The averaged activity detection MAP

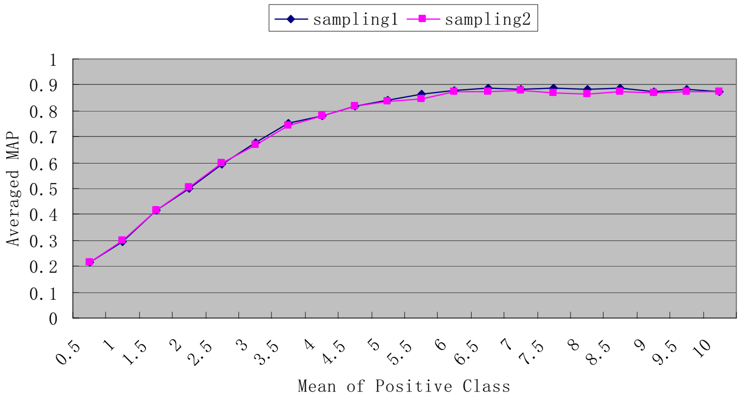


Figure 10: Comparison of two sampling methods (simulated data).

over 20 runs is shown on the y-axis. In Figure 10, two curves of *sampling*1 (half-and-half) and *sampling*2 (odd-and-even) are the performances of two sampling methods. The overlap of two curves shows that there is no significant difference between two sampling methods, especially when $\mu_1 \leq 5.0$, for which the concept detection MAP is relatively low. While μ_1 increases, both of the performances of two samplings increase. When μ_1 is large enough, say, $\mu_1 \geq 6.5$, the concept detection MAP remains at a stable level (nearly perfect as shown in Figure 7), and the curve of *sampling*1 remains higher than that of *sampling*2. This is consist with the comparison using clean data as described in Section 5.2. However, when concept detection is not perfect ($\mu_1 \leq 5.0$), the appearance of concept detection errors will definitely change the underlining concept observation patterns, therefore the two sampling approaches will perform equally. This can be depicted by the overlap of two curves when the value of μ_1 is small.

6.2. Interpreting the Results

From the experimental results we find that our algorithm achieves satisfactory performances in categorizing and detecting various types of daily activities. The experiments were carried out on both perfect concept detection and detection with errors. The application of dynamic concept patterns in activity classification has shown promising result especially for better concept detectors. As shown in Table 4, among the 16 activities investigated, 'Driving', 'Food Shopping', 'Presentation (listen)' and 'Using computer' have the highest accuracy with both precision and recall being 1.00. Other activities like 'Reading', 'Taking bus', 'Using phone', 'Walking' and 'Watching TV' have an F-Score above 0.90.

From Table 4, we find that the highest performances are achieved for activities in which the visual similarity of SenseCam images are high. The stability of concepts decided by image visual features makes it easier to detect these activities. As to the activities involving higher concept diversity, such as 'Child care', 'Cooking', 'Talking', etc. where the subject is likely to be moving around and changing his/her perspective on the environment, the overall accuracies are degraded but still remain at an acceptable level. Only 'Talking' and 'Drinking' have an F-Score lower than 0.80. Note that similar concept dynamics also introduces more misclassifications for activities like 'Drinking' and 'Eating'. In this evaluation, 1 out of 15 'Drinking' samples are detected as 'Eating' while 3 out of 28 'Eating' samples are classified as 'Drinking' activities. From Table 4, we notice that 'Talking' has the lowest recall at 0.65.

This is because 6 of these 17 'Talking' instances are misclassified as 'Drinking' (1 instance), 'General shopping' (1 instance), 'Walking' (3 instances) and 'Child care' (1 instance), due to the very similar concepts like 'Face', 'Hand gesture', etc., which are the cues for 'Talking', but also frequently appear in other activities.

Results also shows the robustness of our algorithm for handling errors in concept detection. Though the performance of activity classification declines with less accurate concept detection (as shown in Figure 7 and Figure 8), the activity detection MAP remains at an acceptable level when more errors are introduced in concept detection, say when $\mu_1 \in [2..4]$. Similar performances are achieved by using both sampling methods as shown in Figure 10 when using erroneous concept detectors. More extreme evaluation is also tested with very poor concept detectors. When μ_1 is assigned very small values, for example, $\mu_1 \leq 1$, the distribution of positive and negative classes are seriously overlapped. In such extreme circumstances, the overall performance of activity classification still drops at a reasonable rate.

7. Conclusions

In this paper we have described a novel application of visual lifelogging where a subject wears a camera that records images of their day-to-day activities, ambiently. Our particular interest is in characterizing the activities and everyday behaviour of the wearer which is distinct from other applications of visual lifelogging like remembrance or re-finding previous events from the past. The novelty of our contribution lies in the fact we have used visual images as the raw source of user observation data, albeit it observation data taken by the rather than of the subject.

Our approach to characterising the wearer's activity profile is to automatically detect the presence or absence of a series of base concepts from the raw lifelong images and to aggregate the appearance of these concepts to indicate higher level activities like eating, shopping, in a meeting or using a phone, which indicate behaviour. The technique we have developed for this is to automatically detect the presence of a set of up to 80 of these base concepts and then to apply a combination of dimensionality reduction to reduce the concept space and k-means clustering, followed by the application of a series of pre-trained HMM models to detect pre-defined specific activities.

Our algorithm was evaluated on a set of SenseCam images and our best results for 16 everyday activities show an F-score which varies from 0.73 to 1.0, with an average of 0.90 across the activities. These results take into account the errors that will inevitably appear in automatic concept detection and the fact that concepts may appear and disappear during an activity or event.

Automatic detection of concepts is already a very active area in multimedia research in general and is applied to video from broadcast TV, movies, TV news etc. as well as to still images. In such work the concepts are quite simple, semantically, mostly corresponding to the appearance of objects (face, person, TV screen, etc.) or characteristics of the environment (indoor, outdoor, sky, etc.). In our work on analysing visual lifelogs, detecting such concepts is useful for helping to find particular events within the lifelong for applications like remembrance. What we have done in this paper is to aggregate the detection of these concepts and from this to infer the appearance of everyday *activities*, corresponding to a higher level of semantics. The overall performance of our technique makes it usable for characterizing the lifestyle and behaviour of subjects. Using the techniques we have presented in this paper we can now explore the lifestyles and behaviors of subjects in visual lifelogging which represents the next step in our future work.

Authors' contributions

Peng Wang carried out the experiments under the supervision of Alan Smeaton. Both authors wrote the paper with equal contribution and both authors approve this submission.

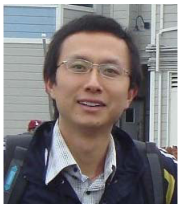
Conflict of interests

The authors have no conflicts of interest in undertaking or reporting this research.

Acknowledgements

This work was supported by Science Foundation Ireland under grant number 07/CE/I1147 and by the China Scholarship Council, neither of which had any involvement in the work, in writing this article or in deciding on its submission to this journal.

Vitae of authors



Peng Wang has a Masters degree in Human-Computer Interaction where he applied infrared technology to muti-touch recognition on tabletop and wall-based computer screens. He developed an ontology of two-handed gestures to recognise a new approach to HCI with low cognitive burden on users. He subsequently completed his PhD dissertation at Dublin City University as part of the CLARITY centre. His research interests are in the area of unobtrusive personal life logging based on context-awareness.

Alan F. Smeaton has been a Professor of Computing at Dublin City University since 1997 and is currently deputy director of CLARITY: Centre for Sensor Web Technologies. He has previously been Dean of Faculty and Head of School and now leads a team of more than 40 researchers in the broad area of the sensor web. The focus of Alan's research work now is in information access for all kinds of human digital memory applications and in exploiting how sensors can be used to help with this.

References

[1] R. Aly, A.R. Doherty, D. Hiemstra, A.F. Smeaton, Beyond shot retrieval: Searching for broadcast news items using language models of concepts, in: Advances in Information Retrieval, volume 5993 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2010, pp. 241–252.

[2] R. Aly, D. Hiemstra, Concept detectors: How good is good enough?, in: Proceedings of the 17th ACM international conference on Multimedia, MM '09, ACM, New York, NY, USA, 2009, pp. 233–242.

[3] R. Aly, D. Hiemstra, F. de Jong, P. Apers, Simulating the future of concept-based video retrieval under improved detector performance, Multimedia Tools and Applications (2011) 1–29.

[4] E. Berry, N. Kapur, L. Williams, S. Hodges, P. Watson, G. Smyth, J. Srinivasan, R. Smith, B. Wilson, K. Wood, 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 17 (2007) 582–601.

[5] G. Browne, E. Berry, N. Kapur, S. Hodges, G. Smyth, P. Watson, K. Wood, Sensecam improves memory for recent events and quality of life in a patient with memory retrieval difficulties, Memory 19 (2011) 713–722.

[6] R. Chilvers, S. Corr, S. Hayley, Investigation into the occupational lives of healthy older people through their use of time, Australian Occupational Therapy Journal 57 (2010) 24–33.

[7] C.O. Conaire, D. Connaghan, P. Kelly, N.E. O'Connor, M. Gaffney, J. Buckley, Combining inertial and visual sensing for human action recognition in tennis, in: Proceedings of the first ACM international workshop on analysis and retrieval of tracked events and motion in imagery streams, ARTEMIS '10, ACM, New York, NY, USA, 2010, pp. 51–56.

[8] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science 41 (1990) 391–407.

[9] A.R. Doherty, N. Caprani, C. O Conaire, V. Kalnikaite, C. Gurrin, N.E. O'Connor, A.F. Smeaton, Passively recognising human activities through lifelogging, Computers in Human Behavior 27 (2011) 1948–1958.

[10] S. Ebadollahi, L. Xie, S.F. Chang, J.R. Smith, Visual event detection using multi-dimensional concept dynamics., in: ICME, IEEE, 2006, pp. 881–884.

[11] P.A. Gambrel, R. Cianci, Maslow's hierarchy of needs: Does it apply in a collectivist culture, Journal of Applied Management and Entrepreneurship 8 (2003) 143–161.

[12] N. Harte, D. Lennon, A. Kokaram, On parsing visual sequences with the Hidden Markov Model, J. Image Video Process. 2009 (2009) 6:1–6:13.

[13] M. Hill, G. Hua, A. Natsev, J.R. Smith, L. Xie, B. Huang, M. Merler, H. Ouyang, M. Zhou, IBM Rsesearch TRECVid-2010 video copy detection and multimedia event detection system, in: NIST TRECVid Workshop.

[14] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, K. Wood, SenseCam: A retrospective memory aid, in: Proc. 8th International Conference on Ubicomp, Orange County, CA, USA, pp. 177–193.

[15] B. Huurnink, K. Hofmann, M. de Rijke, Assessing concept selection for video retrieval, in: MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval, ACM, New York, NY, USA, 2008, pp. 459–466.

[16] Y.G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, S.F. Chang, Columbia-UCF TRECVid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching, in: NIST TRECVid Workshop.

[17] C.H. Kaczkowski, P.J.H. Jones, J. Feng, H.S. Bayley, Four-day multimedia diet records underestimate energy needs in middle-aged and elderly women as determined by doubly-labeled water., Journal of Nutrition 130 (2000) 802–5.

[18] D. Kahneman, A.B. Krueger, D.A. Schkade, N. Schwarz, A.A. Stone, A survey method for characterizing daily life experience: The day reconstruction method, Science 306 (2004) 1776–1780.

[19] S. Karaman, J. Benois-Pineau, R. Megret, J. Pinquier, Y. Gaestel, J.F. Dartigues, Activities of daily living indexing by hierarchical HMM for dementia diagnostics, in: The 9th International Workshop on Content-Based Multimedia Indexing (CBMI), Madrid, Spain, pp. 79–84.

[20] P. Kelly, A.R. Doherty, E. Berry, S. Hodges, A.M. Batterham, C. Foster, Can we use digital life-log images to investigate active and sedentary travel behaviour? Results from a pilot study, International Journal of Behavioral Nutrition and Physical Activity 8 (2011) 44–.

[21] D.T. Kenrick, V. Griskevicius, S.L. Neuberg, M. Schaller, Renovating the pyramid of needs: Contemporary extensions built upon ancient foundations, Perspectives on Psychological Science 5 (2010) 292–314.

[22] G. Lakoff, Women, Fire, and Dangerous Things, University of Chicago Press, 1990.

[23] T.K. Landauer, P.W. Foltz, D. Laham, An Introduction to Latent Semantic Analysis, Discourse Processes (1998) 259–284.

[24] M. Law, S. Steinwender, L. Leclair, Occupation, health and well-being, Canadian Journal of Occupational Therapy 65 (1998) 81–91.

[25] M.L. Lee, A.K. Dey, Lifelogging memory appliance for people with episodic memory impairment, in: Proceedings of the 10th international conference on Ubiquitous computing, UbiComp '08, ACM, New York, NY, USA, 2008, pp. 44–53.

[26] W.H. Lin, A.G. Hauptmann, Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts, in: IEEE International Conference on Multimedia and Expo, IEEE Computer Society, Los Alamitos, CA, USA, 2006, pp. 41–44.

[27] X. Ma, D. Schonfeld, A.A. Khokhar, Video event classification and image segmentation based on non-causal multi-dimensional hidden markov models, Transactions on Image Processing. 18 (2009) 1304–1313.

[28] A.H. Maslow, A theory of human motivation, Psychological Review 50 (1943) 370–396.

[29] K. McKenna, K. Broome, J. Liddle, 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 54 (2007) 273–284.

[30] R. Mégret, V. Dovgalecs, H. Wannous, S. Karaman, J. Benois-Pineau, E.E. Khoury, J. Pinquier, P. Joly, R. André-Obrecht, Y. Gaëstel, J.F. Dartigues, The IMMED project: wearable video monitoring of people with age dementia, in: Proceedings of the international conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 1299–1302.

[31] W. Mittelman, Maslow's study of self-actualization - a reinterpretation, Journal of Humanistic Psychology 31 (1991) 114–135.

[32] M. Naphade, J.R. Smith, J. Tesic, S.F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, Large-scale concept ontology for multimedia, IEEE Multimedia 13 (2006) 86–91.

[33] M.R. Naphade, L. Kennedy, J.R. Kender, S.F. Chang, J.R. Smith, P. Over, A. Hauptmann, A light scale concept ontology for multimedia understanding for TRECVid 2005, Technical Report, IBM Research, 2005.

[34] A. Nijholt, Google home: Experience, support and re-experience of social home activities, Information Sciences 178 (2008) 612 – 630.

[35] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, in: Readings in speech recognition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, pp. 267–296.

[36] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, M. Hansen, 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, ACM Press, Cork, Ireland, 2007, pp. 13–17.

[37] M. Ros, M. Cuéllar, M. Delgado, A. Vila, Online recognition of human activities and adaptation to habit changes by means of learning automata and fuzzy temporal windows, Information Sciences 220 (2013) 86 – 101.

[38] A.F. Smeaton, P. Over, W. Kraaij, Evaluation campaigns and TRECVid, in: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, ACM Press, New York, NY, USA, 2006, pp. 321–330.

[39] A.F. Smeaton, P. Over, W. Kraaij, High-level feature detection from video in TRECVid: a 5-year retrospective of achievements, in: A. Divakaran (Ed.), Multimedia Content Analysis, Theory and Applications, Springer Verlag, Berlin, 2009, pp. 151–174.

[40] C.G.M. Snoek, M. Worring, Concept-based video retrieval, Foundations and Trends in Information Retrieval 2 (2009) 215–322.

[41] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.M. Geusebroek, A.W.M. Smeulders, The challenge problem for automated detection of 101 semantic concepts in multimedia, in: Proceedings of the 14th annual ACM international conference on Multimedia, MULTIMEDIA '06, ACM, New York, NY, USA, 2006, pp. 421–430.

[42] P.L. St. Jacques, M.A. Conway, M.W. Lowder, , R. Cabeza, Watching my mind unfold versus yours: An fMRI study using a novel camera technology to examine neural differences in self-projection of self versus other perspectives, Journal of Cognitive Neuroscience 23 (2011) 1275–1284.

[43] T.L.P. Tang, A.H. Safwat Ibrahim, Importance of human needs during retrospective peacetime and the persian gulf war: Mideastern employees, International Journal of Stress Management 5 (1998) 25–37.

[44] T.L.P. Tang, W.B. Safwat Ibrahim, Abdul H.and West, Effects of war-related stress on the satisfaction of human needs: The united states and the middle east, International Journal of Management Theory and Practices 3 (2002) 35–53.

[45] P. Toharia, O. Robles, A.F. Smeaton, A. Rodrguez, Measuring the influence of concept detection on video retrieval, in: X. Jiang, N. Petkov

(Eds.), Computer Analysis of Images and Patterns, volume 5702 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2009, pp. 581–589.

[46] UK Government, Office for National Statistics (ONS), The United Kingdom 2000 time use survey, Available from `Availablefromhttp://www.ons.gov.uk/ons/rel/lifestyles/time-use/2000-edition/time-use-survey--technical-report.pdf`, 2003. Last accessed: January, 2012.

[47] UK Government, Office for National Statistics (ONS), The Time Use Survey, 2005: How We Spend Our Time (Amended), Available from `http://http://www.ons.gov.uk/ons/rel/lifestyles/time-use/2005-edition/time-use-survey-2005--how-we-spend-our-time.pdf`, 2006. Last accessed: January 2012.

[48] A.V. Vasilakos, L. Wei, T.H.D. Nguyen, T.C.T. Qui, L.C. Chen, C. Boj, D. Diaz, A.D. Cheok, G. Marentakis, Interactive theatre via mixed reality and ambient intelligence, Information Sciences 178 (2008) 679 – 693.

[49] P. Wang, A.F. Smeaton, Semantics-based selection of everyday concepts in visual lifelogging, International Journal of Multimedia Information Retrieval 1 (2012) 87–101. 10.1007/s13735-012-0010-8.

[50] L. Xie, S.F. Chang, A. Divakaran, H. Sun, Unsupervised mining of statistical temporal structures in video, in: A. Rosenfeld, D. Doremann, D. Dementhon (Eds.), Video Mining, Kluwer Academic Publishers, 2003.

[51] J.M. Zacks, T.S. Braver, M.A. Sheridan, D.I. Donaldson, A.Z. Snyder, J.M. Ollinger, R.L. Buckner, M.E. Raichle, Human brain activity time-locked to perceptual event boundaries., Nature Neuroscience 4 (2001) 651–655.