Validating the Detection of Everyday Concepts in Visual Lifelogs

Daragh Byrne[1,2], Aiden R. Doherty[1,2], Cees G.M. Snoek[3], Gareth G.F. Jones[1], and Alan F. Smeaton[1,2]

[1] Centre for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, Ireland
[2]CLARITY: Centre for Sensor Web Technologies
{daragh.byrne, aiden.doherty, gareth.jones, alan.smeaton}@computing.dcu.ie
[3] ISLA, University of Amsterdam, Kruislaan 403, 1098SJ Amsterdam, The Netherlands
cgmsnoek@ uva.nl

Abstract. The Microsoft SenseCam is a small lightweight wearable camera used to passively capture photos and other sensor readings from a user's day-to-day activities. It can capture up to 3,000 images per day, equating to almost 1 million images per year. It is used to aid memory by creating a personal multimedia lifelog, or visual recording of the wearer's life. However the sheer volume of image data captured within a visual lifelog creates a number of challenges, particularly for locating relevant content. Within this work, we explore the applicability of semantic concept detection, a method often used within video retrieval, on the novel domain of visual lifelogs. A concept detector models the correspondence between low-level visual features and high-level semantic concepts (such as indoors, outdoors, people, buildings, etc.) using supervised machine learning. By doing so it determines the probability of a concept's presence. We apply detection of 27 everyday semantic concepts on a lifelog collection composed of 257,518 SenseCam images from 5 users. The results were then evaluated on a subset of 95,907 images, to determine the precision for detection of each semantic concept and to draw some interesting inferences on the lifestyles of those 5 users. We additionally present future applications of concept detection within the domain of lifelogging.

Keywords: Microsoft SenseCam, lifelog, passive photos, concept detection, supervised learning

1 Introduction

Recording of personal life experiences through digital technology is a phenomenon we are increasingly familiar with: music players, such as iTunes, remember the music we listen to frequently; our web activity is recorded in web browsers' "History"; and we capture important moments in our life-time through photos and video [1]. This concept of digitally capturing our memories is known as lifelogging. Lifelogging and memory capture was originally envisaged to fulfill at least part of Vannevar Bush's 1945 MEMEX vision. Bush describes his MEMEX as a collection in which a person could store all of their life experience information including photographs, documents

and communications "*and which is mechanized so that it may be consulted with exceeding speed and flexibility*." [3].

A visual lifelogging device, such as the SenseCam, will capture approximately 1 million images per year. The sheer volume of photos collected, and the rate at which a collection can grow, pose significant challenges for the access, management, and utility of such a lifelog. However, inroads to resolving some of the concerns relating to these issues have already been made. For example, in prior work we proposed the aggregation of individual images within a visual lifelog into higher level discrete 'events' which represent single activities in a user's day [8]. Furthermore, work has been carried out to investigate how best to select a single representative keyframe image which best summarises a given event [7]. Lee *et. al.* have constructed an event-oriented browser which enables a user to browse each day in their collection through a calendar controlled interface [19]. This interface allows the 'gisting' or recap of an entire day's activities by presenting a visual summary of the day. The benefit of such daily summaries has been highlighted in the results of a preliminary study carried out between Microsoft Research and Addenbrooke's Hospital, Cambridge, U.K. where visual lifelog recordings notably improved subjects' recall of memories [15].

A fundamental requirement outlined in Bush's MEMEX [3] is that we must provide on-demand, rapid and easy access to the memories and experiences of interest and to achieve this we must be able to support high quality retrieval. While many steps have been taken towards managing such an ever-growing collection [7,8,20], we are still far from achieving Bush's original vision. This is mainly due to the fact that we cannot yet provide rapid, flexible access to content of interest from the collection.

The most obvious form of content retrieval is to offer refinement of the lifelog collection based on temporal information. Retrieval may also be enabled based on the low-level visual features of a query image. However, in order for such a search to be effective the user must provide a visual example of the content they seek to retrieve and there may be times when a user will not possess such an example, or that it may be buried deep within the collection. Augmentation and annotation of the collection with sources of context metadata is another method by which visual lifelogs may be made searchable. Using sources of context such as location or weather conditions has been demonstrated to be effective in this regard [4,10]. There are, however, limitations to these approaches as well, most importantly any portion of the collection without associated context metadata would not be searchable. Moreover, while information derived from sensors such as Bluetooth and GPS [4] may cover the 'who' and the 'where' of events in an individual's lifelog, however, they do not allow for the retrieval of relevant content based on the 'what' of an event.

An understanding of the 'what' or the semantics of an event would be invaluable within the search process and would empower a user to rapidly locate relevant content. Typically, such searching is enabled in image tools like Flickr through manual user contributed annotations or 'tags', which are then used to retrieve visual content. Despite being effective for retrieval, such a manual process could not be practical within the domain of lifelogging, since it would be far too time and resource intensive given the volume of the collection and the rate at which it grows. Therefore we should explore methods for automatic annotation of visual lifelog collections.

One such method is concept detection, an often employed approach in video retrieval [22,24,27], which aims to describe visual content with confidence values

indicating the presence or absence of object and scene categories. Although it is hard to bridge the gap between low-level features that one can extract from visual data and the high-level conceptual interpretation a user gives to this data, the video retrieval field has made substantial progress by moving from specific single concept detection methods to generic approaches. Such generic concept detection approaches are achieved by fusion of color-, texture-, and shape-invariant features [11,12,14,25], combined with supervised machine learning using support vector machines [5,26]. The emphasis on generic indexing by learning has opened up the possibility of moving to larger concept detector sets [16,23,28]. Unfortunately these concept detector sets are optimized for the (broadcast) video domain only, and their applicability to other domains such as visual lifelog collections remains as of yet unclear.

Visual lifelog data, and in particular Microsoft SenseCam data – the source for our investigation - is markedly different from typical video or photographic data and as such presents a significantly more challenging domain for visual analysis. SenseCam images tend to be of low quality owing to: their lower visual resolution; their use of a fisheye lens which distorts the image somewhat but increases the field of view; and a lack of flash resulting in many images being much darker than desired for optimal visual analysis. Also, almost half of the images are generally found to contain non-desirable artifacts such as grain, noise, blurring or light saturation [13]. Thus our investigation into the precision and reliability of semantic concept detection methods will provide important insights into their application for visual lifelogs.

The rest of this paper is organised as follows: Section 2 details how we apply concept detection to images captured by the SenseCam lifelogging device; section 3 quantitatively describes how accurate our model is in detecting concepts; section 4 provides interesting inferences on the lifestyles of our users using the detected concepts; while sections 5 and 6 finally summarise this work and detail many interesting future endeavours to be investigated.

2. Concept Detection Requirements in the Visual Lifelog Domain

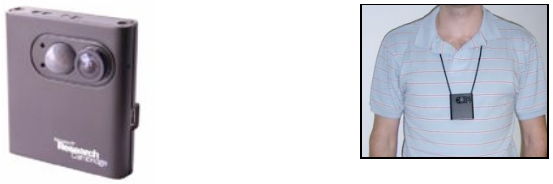
The major requirements for semantic concept detection on visual lifelogs are as follows: 1) Identification of Everyday Concepts; 2) Reliable and Accurate Detection; and 3) Identification of Positive and Negative Examples. We now discuss how we followed these steps with respect to lifelog images captured by a SenseCam.

Use Case: Concept Detection in SenseCam Images

To study the applicability of concept detection in the lifelog domain we make use of a device known as the SenseCam. Microsoft Research in Cambridge, UK, have developed the SenseCam as a small wearable device that passively captures a person's day-to-day activities as a series of photographs and readings from in-built sensors [15]. It is typically hung from a lanyard around the neck and, so is oriented towards the majority of activities which the user is engaged in. Anything in the view of the wearer can be captured by the SenseCam because of its fisheye lens. At a minimum the SenseCam will automatically take a new image approximately every 50 seconds,

but sudden changes in the environment of the wearer, detected by onboard sensors, can trigger more frequent photo capture. The SenseCam can take an average of 3,000 images in a typical day and, as a result, a wearer can very quickly build large and rich photo collections. Already within a year, the lifelog photoset will grow to approximately 1 million images!

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



2.1 Collection Overview

In order to appropriately evaluate concept detection we organised the collection of a large and diverse dataset of 257,518 SenseCam images gathered by five individual users. In order to further ensure diversity, there was no overlap between the periods captured within each user's dataset. A breakdown of the collection is illustrated in Table 1. It is worth noting that not all collections featured the same surroundings. Often collections were subject to large changes in location, behaviour, and environments. This allowed us to more reliably determine the robustness of concept detection in this domain.

User	Total Images	Number of Concepts Annotated	Days Covered
1	79,595	2,180	35
2	76,023	9,436	48
3	42,700	27,223	21
4	40,715	28,023	25
5	18,485	11,408	8

Table 1. An overview of the image collection used.

2.2 Determining LifeLog Concepts

Current approaches to semantic concept detection are based on a set of positive and a set of negative labeled image examples by which a classifier system can be trained (see section 2.3). Consequently, as part of this investigation we identify the concepts within the collection for which a set of training examples would be collected. In order to determine the everyday concepts within the collection, a subset of each user's collection was visually inspected by playing the images sequentially at highly accelerated speed. A list of concepts previously used in video retrieval [22,23] and agreed upon as applicable to a SenseCam collection were used as a starting point. As

Concept / User	1	2	3	4	5	All
Total to Annotate	16111	14787	8593	8208	3697	51396
Indoors	1093	1439	6790	6485	3480	19287
Hands	1	17	4727	3502	2402	10649
Screen (computer/laptop)	7	1101	4699	2628	2166	10601
Office	7	78	4759	2603	336	7783
People	0	1775	573	3396	889	6633
Outdoors	250	915	1248	812	67	3292
Faces	0	553	101	1702	662	3018
Meeting	0	808	0	1233	355	2396
Inside of vehicle, not driving (airplane, car, bus)	257	1326	420	223	0	2226
Food (eating)	0	795	349	870	129	2143
Buildings	140	49	981	621	62	1853
Sky	0	202	720	525	66	1513
Road	125	0	231	648	4	1008
Tree	24	44	378	469	42	957
Newspaper/Book (reading)	0	85	13	520	309	927
Vegetation	0	3	255	468	52	778
Door	28	0	279	128	144	579
Vehicles (external view)	33	0	322	121	4	480
Grass	0	122	99	190	33	444
Holding a cup/glass	0	0	21	353	44	418
Giving Presentation / Teaching	0	43	0	309	0	352
Holding a mobile phone	0	4	54	28	147	233
Shopping	0	75	102	48	3	228
Steering wheel (driving)	208	0	0	0	0	208
Toilet/Bathroom	6	0	75	93	0	174
Staircase	0	2	26	48	11	87
View of Horizon	1	0	1	0	1	3

Table 2. An outline of the 27 concepts and the no. of positive examples per concept and per user.

a new identifiable 'concept' was uncovered within the collection it was added to this list. Each observed repetition of the concept gave it additional weight and ranked it more highly for inclusion. Over 150 common concepts were identified in this process. It was decided that the most representative (i.e. everyday) concepts should be selected and as such these were then narrowed to just 27 core concepts through iterative review and refinement. Criteria for this refinement included the generalisability of the concept across collections and users. For example, the concepts 'mountain' and 'snow' occurred in User 1's collection frequently but could not be considered as an

everyday concept as it was not present in the remaining collections. As such the 27 concepts represent a set of everyday core concepts most likely to be collection independent, which should consequently be robust with respect to the user and setting. These core concepts are outlined in Figure 2 using visual examples from the collection. Given that some concepts are related (e.g. it is logical to expect that 'buildings' and 'outdoors' would co-occur), it is important to note that each image may contain multiple (often semantically related) concepts.

A large-scale manual annotation activity was undertaken to provide the required positive and negative labeled image examples. As annotating the entire collection was impractical and given that SenseCam images tend to be temporally consistent the collection was skimmed by taking every 5th image. Collection owners annotated their own SenseCam images for the presence of each of the concepts. As by their nature lifelog images are highly personal, it is important for privacy reasons that it is only the owner of the lifelog images who labels his or her images. Therefore, collection owners annotated their own SenseCam images for each concept. This also provided the opportunity for them to remove any portion of their collection they did not wish to have included as part of this study. All users covered their entire skimmed collection with the exception of User 1, who only partially completed the annotation process on a subset of his collection. The number of positive examples for each concept and for each user is presented in Table 2.

2.3 Concept Detection Process

Our everyday concept detection process is composed of three stages: 1) supervised learning, 2) visual feature extraction, and 3) feature and classifier fusion, each of these stage uses the implementation detailed below.

Supervised Learner: We perceive concept detection in lifelogs as a pattern recognition problem. Given pattern x, part of an image i, the aim is to obtain a probability measure, which indicates whether semantic concept ω_j is present in image i. Similar to [16,24,27,28], we use the Support Vector Machine (SVM) framework [26] for supervised learning of concepts. Here we use the LIBSVM implementation [5] with radial basis function and probabilistic output [21]. We obtain good SVM settings by using an iterative search on a large number of parameter combinations.

Visual Feature Extraction: For visual feature extraction we adopt the well-known codebook model, see e.g. [17], which represents an image as a distribution over codewords. We follow [25] to build this distribution by dividing an image in several overlapping rectangular regions. We employ two visual feature extraction methods to obtain two separate codebook models, namely: 1) *Wiccest features*, which rely on natural image statistics and are therefore well suited to detect natural sceneries, and 2) *Gabor features*, which are sensitive to regular textures and color planes, and therefore well suited for the detection of man-made structures. Both these image features measure colored texture.

Wiccest features [11] utilise natural image statistics to model texture information. Texture is described by the distribution of edges in a certain image region. Hence, a histogram of a Gaussian derivative filter is used to represent the edge statistics. It was shown in [12] that the complete range of image statistics in natural textures can be well modeled with an integrated Weibull distribution, which in turn can be

Figure 2. Visual examples of each of the 27 everyday concepts as detected and validated for the lifelog domain.

characterised by just 2 parameters. Thus, 2 Weibull parameter values for the *x*-edges and *y*-edges of the three color channels yields a 12 dimensional descriptor. We construct a codebook model from this low-level region description by computing the similarity between each region and a set of 15 predefined semantic color-texture

patches (including e.g. sand, brick, and water), using the accumulated fraction between their Weibull parameters as a similarity measure [25]. We perform this procedure for two region segmentations, two scales, the *x*- and the *y*-derivatives, yielding a codebook feature vector of 120 elements we term *w*.

Gabor filters may be used to measure perceptual surface texture in an image [2]. Specifically, Gabor filters respond to regular patterns in a given orientation on a given scale and frequency. In order to obtain an image region descriptor with Gabor filters we follow these three steps: 1) parameterise the Gabor filters, 2) incorporate color invariance, and 3) construct a histogram. First, the parameters of a Gabor filter consist of orientation, scale and frequency. We use four orientations, 0º, 45º, 90º, 135º, and two (scale, frequency) pairs: (2.828, 0.720), (1.414, 2.094). Second, color responses are measured by filtering each color channel with a Gabor filter. The W color invariant is obtained by normalising each Gabor filtered color channel by the intensity [14]. Finally, a histogram is constructed for each Gabor filtered color channel. We construct a codebook model from this low-level region description by again computing the similarity between each region and a set of 15 predefined semantic color-texture patches, where we use histogram intersection as the similarity measure. Similar to the procedure for *w*, this yields a codebook feature vector of 120 elements we term *g*.

Feature and Classifier Fusion: As the visual features *w* and *g* emphasise different visual properties, we consider them independent. Hence, much is to be expected from their fusion. We employ fusion both at the feature level as well as the classifier level. Although the vectors *w* and *g* rely on different low-level feature spaces, their codebook model is defined in the same codeword space. Hence, for feature fusion we can concatenate the vectors *w* and *g* without the need to use normalisation or transformation methods. This concatenation yields feature vector *f* .

For each of the feature vectors in the set $\{w, g, f\}$ we learn a supervised classifier. Thus for a given image *i* and a concept ω_j , we obtain three probabilities, namely: $p(\omega_j | w_i)$, $p(\omega_j | g_i)$, and $p(\omega_j | f_i)$, based on the same set of labeled examples. To maximize the impact of our labeled examples, we do not rely on supervised learning in the classifier fusion stage. Instead, we employ average fusion of classifier probability scores, as used in many visual concept detection methods [16,24,27,28]. After classifier fusion we obtain our final concept detection score, which we denote $p(\omega_{ij})$.

3 Validation of Everyday Concept Detection

In order to validate $p(\omega_{ij})$, we manually judged a subset of the collection. To make a determination of its presence we employ a thresholding technique which divides the collection into those considered to contain the concept and those which do not. To achieve this, while simultaneously selecting a threshold value for each concept, we use the Kapur automatic thresholding technique [18]. Since this entropy based non-parametric method does not require any training, it can be easily applied to such a broad collection. We consider any images above the threshold value to be positive examples of that concept. Similarly, any frames below the threshold were considered

Concept Name	No. Samples Provided	Number of Judgements	System Positive Accuracy	System Negative Accuracy
Indoor	19,287	3,271	82%	45%
Sky	1,513	4,099	79%	90%
Screen	10,601	3,761	78%	85%
Shopping	228	3,500	75%	99%
Office	7,783	3,436	72%	77%
steeringWheel	208	3,936	72%	99%
Door	579	3,512	69%	86%
Hands	10,649	3,399	68%	68%
Veg	778	3,336	64%	97%
Tree	957	3,736	63%	98%
Outdoor	3,292	3,807	62%	97%
Face	3,018	3,452	61%	91%
Grass	444	3,765	61%	99%
insideVehicle	2,226	3,604	60%	93%
Buildings	1,853	3,654	59%	98%
Reading	927	3,420	58%	94%
Toilet	174	3,683	58%	99%
Stairs	87	2,927	48%	100%
Road	1,008	3,548	47%	96%
vehiclesExternal	480	3,851	46%	98%
People	6,633	3,024	45%	90%
Eating	2,143	3,530	41%	97%
holdingPhone	233	3,570	39%	99%
holdingCup	418	3,605	35%	99%
Meeting	2,396	3,534	34%	94%
presentation	352	3779	29%	99%
viewHorizon	3	3168	23%	98%

Table 4. Accuracy of detection for each concept (Sorted by 'System Positive Accuracy').

as negative. Next, nine participants manually judged a subset of system positive and negative examples for each concept. In order to judge the intercoder reliability, consistency and accuracy of each annotator's performance; 50 positive and 50 negative examples per concept were randomly selected for judgment by each of the 9 annotators. Additionally, per concept, another 150 system judged positive and negative frames were randomly selected and assigned to every annotator. This resulted in almost 1400 positive and negative unique images per concept to be judged by the 9 annotators (50 to be judged by all 9 plus 9x150 individual judgments).

To support this judgment process a custom annotation tool was developed. Participants were presented with a tiled list of images and given instructions on how to appropriately judge them against each concept. Users simply clicked an image to mark it as a positive match to the provided concept. For each concept both system judged positive and negative images were presented in tandem and were randomly selected from the total pool of judgments to be made. Annotating in this fashion allowed a total of 95,907 judgments made across all users on 70,659 unique concept validation judgments (which used 58,785 unique images). This yielded a detailed validation of both the images considered positive and negative for each concept.

Each annotator provided judgments for a shared set of 100 images per concept. These images were then used to determine the amount of agreement in the judgments among the nine annotators. An understanding of this 'intercoder reliability' is important as it validates the reliability of the overall annotation process and the performance of the annotators in general. This allows us to ensure that the outcome of the validation process is wholly reliable. The intercoder reliability was determined to be 0.68 for all judgments completed using Fleiss's Kappa [9]. As such the annotations provided by these participants are consistent and have very good inter-coder agreement. Examination at the concept level shows 18 of the 27 concepts had at least 0.6 agreement which is substantial according to Landis and Koch [19]. While examination of individual concepts reveals some variability in inter-rater reliability and a number of lower than anticipated agreement for a minority of the concepts (k=0.64 average overall; minimum 0.37 – view of horizon; maximum 0.86 – steering wheel), given that the number of judgments made per annotator was extremely large, this may have had the affect of reducing the overall magnitude of the value. We believe that the agreement between the annotators is sufficiently reliable to use these judgments to validate the automatically detected concepts.

3.1 Analysis of System Results

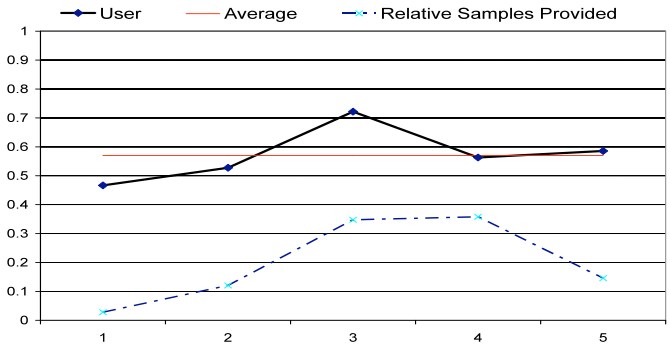
From the 95,907 judged results, 72,143 (75%) were determined to be correctly classified by the system. This figure, however, includes both positive and negative images for a concept as determined by the system. Of all those judgments, the system correctly identified 57% of true positives overall. 93% of system negatives were correct, meaning that only 7% of true positives were missed across all the concepts on the entire dataset.

Given the variation in complexity of the concepts and in the level of semantic knowledge they attempt to extract, it is unsurprising that there is notable differences in their performance and accuracy. Furthermore, the quality, variance and number of training examples will impact on the performance of an individual detector and as such these may be factors in their differing performances. This is outlined in Table 4 which is ordered by concept performance. From this it is clear that the 'indoor' detector worked best, with several other concepts providing similarly high degrees of accuracy. These include the "steeringWheel", "office", "shopping", and "screen" concepts. It is also interesting to note from Table 4, that with the exception of the 'indoor' concept, there are very few missed true positive examples in our large set of

judged images. As the images were collected from 5 separate users it is interesting to explore the degree of variance in the performance between concepts (in terms of true positives). The performance ranged from 46% to 72%, but as illustrated in Fig 3, the deviation of results is not so large when the number of concept training samples provided to the system is considered (the blue dashed line at the bottom of Figure 3). There exists a strong correlation of 0.75 between the number of examples provided by each user to the system and the actual system classification results on the set of 95,907 judged results.

17 of the 27 concepts are at least 58% accurate in correctly identifying positive image examples for a given concept. Apart from the "people" concept we argue that the performance of the other concepts can be improved by providing more positive labeled image examples for each concept. We believe the concept detection results on SenseCam images are sufficiently reliable such that inferences on user patterns and behaviour may be made.

Figure 3. Performance of all concepts on users' collections.



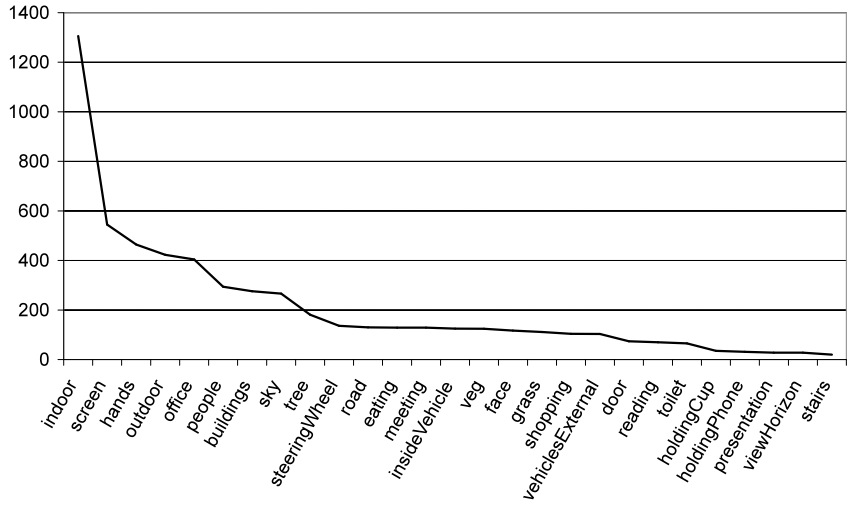
4 Event-Based Results Of Concept Detection Activity

The concept detection results provided for the 219,312 images across the five content owners were then further analysed and investigated. There was wide variance in the number of images determined to be relevant from concept to concept. For example, 107,555 "indoor" images were detected, while just 72 images were detected as being of the "viewHorizon" concept. A number of concepts have a semantic relationship. For example, the "tree" (5,438) and "vegetation" (4,242) concepts closely relate to one another and as such have a similar number of positive examples. However, conversely, within the collection there were many more images containing "people" (29,794) than "face" (11,516) concepts. This is initially a little counterintuitive. While this may be attributed to the 'people' detector being relatively unreliable, there is another more probable explanation. Often a wearer will be for example on the street walking, or on a bus, and faces will not be clearly identifiable either as a result of people facing away from the wearer or being in the distance.

All of the collection owners are researchers and it was also noticed that the concepts with the highest number of occurrences closely match that of what would be expected for such users, e.g. “indoor”, “screen”, “hands” (e.g. on keyboard), “office”, “meeting”, etc. It should be noted that the concept detectors returns results that quite accurately fit those concepts that our users most commonly encounter.

As previously mentioned, using techniques as outlined in [8], a collection of SenseCam images can be aggregated into a higher level discrete unit known as an ‘event’. This has the function of reducing the approximate 3,000 images captured in an average day to on average 20 ‘events’, making the collection far more manageable for its owner. It also has the added advantage of more closely approximating the ‘episodic’ units in which we as humans consider our past experiences. While our analysis has focused on the accuracy of the concept detectors at the image level, it is, as such, worth considering the concepts as detected at the higher level of ‘events’.

Figure 6. Number of "events" per concept

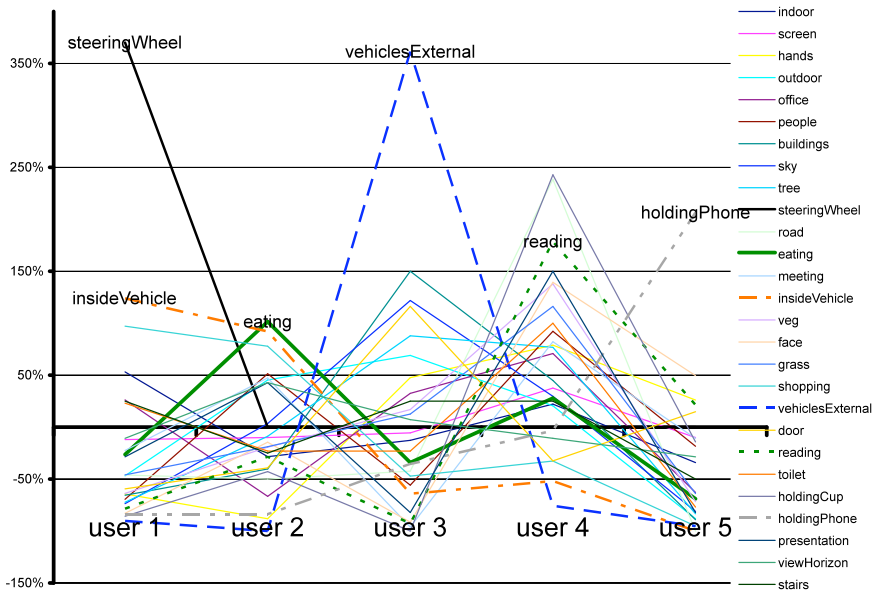


Using this event segmentation, we identified a total of 3,030 events in the collection. As a concept will likely be related to the higher-level activity embodied by an event, the concepts were further explored at the event level. To achieve this we determined the concentration of the concepts within each event, e.g. a ‘working at the computer’ event will have a high percentage of images containing the “screen” and “indoor” concept. To be consistent with our prior analysis on the image-based level, we again made use of the Kapur thresholding approach [18] to determine whether an event had a sufficiently dense concentration of images of a particular concept. That is to say, for each concept we determined the percentage of images in an event that must contain that concept, in order for that event to be considered to represent the concept. “Building events”, for example, should have at least 28% of their images identified as being “building”, while “indoor” events should have at least 48% of images with the “indoor” concept, etc. It is evident from observation of Fig. 6, the event and image level are very similar in their distribution of concepts. As expected the “indoor”, “screen”, and “office” concepts are still very common when they are considered in

terms of events. Likewise, there is a very small sample of events that are under the concept types of “stairs”, “viewHorizon”, “holdingPhone”, etc.

It is particularly interesting to consider, as illustrated by Fig. 7, the number of events each user had (relative to the size of their collection) for each concept type. To explain further, for user 1 the “steeringWheel” concept occurred over 350% more frequently than the median of all the other users. The median value is the x-axis, i.e. 0% different to the median! As such, this graph gives an outline of the differing lifestyles of the users.

Figure 7. Deviation of user examples to the median (per concept)



For user 1 it is interesting to observe that he has much more “steeringWheel” and “insideVehicle” events than the other users. This is indeed to be expected given that this user is the only one of our collectors who regularly drove a car. In fact in providing the initial set of 208 positive examples of this concept, user 1 was responsible for all of these images.

For user 2 it is noticeable that there are relatively many more “eating” events. An explanation for this is that this user wore the SenseCam to quite a few conferences, which included quite a few meals. Also this user was generally quite diligent in wearing his SenseCam for breakfast and supper. It is interesting to note that this user did not provide the most samples for this concept detector to train on initially.

For user 3 there were many more “vehiclesExternal” events than for the other users. We attribute this to the fact that this user provided 67% of the samples for the concept detector to train on. While with user number 4, it is quite evident that he has many more “reading” events. An explanation is that this researcher is very diligent in terms of reading his literature, and is well known for this trait.

User number 5 seemingly had an unusually high number of "holdingPhone" events. We explain this by the fact that this user was conducting experiments with his mobile phone at the time of capturing these SenseCam images. Due to the nature of the experiment he was additionally capturing surrogate image data using the mobile phone's camera and as such was carrying the phone throughout the data collection period. As a result many of his events (relative to the other users) were annotated as being examples of containing the "holdingPhone" concept. Also this user provided 63% of the samples for this particular concept to train on.

5 Future Work

This study was designed to investigate the feasibility of applying automatic concept detection methods in the domain of visual lifelogs. With the reliability of such techniques now validated, a number of explorations still remain.

First, there is a great deal of scope to enhance the robustness of such approaches. For the most part, frames which compose an event tend to be temporally consistent in their visual properties and in the concepts they contain. There is potential to leverage this property to further validate the presence of a concept. In addition to the photos the SenseCam captures, it also continually records the readings from its onboard sensors (light, temperature, accelerometer). The measurements taken from these sensors could be useful to augment and enhance the detection of the concepts from visual features or to detect wholly new 'activity-centric' concepts as in [6]. Other contextual sources such as Bluetooth and GPS could also be used in augmentation [4].

Concept based retrieval has been extremely effective in the domain of digital video [24]. As such retrieval using automatically detected concepts within visual lifelogs should be explored. The performance and utility of such concept-based retrieval approaches should be compared with other methods such as using social context [4].

Finally, exploration into semi-automatic concept annotation of a collection could be achieved through active learning. This would offer the ability to create and train new concept detectors as users explore and annotate their collections. By enabling efficient automatic annotation of new content, while providing flexibility to users to personalise and extend the set of concepts, further utility would be added to lifelogs.

6 Conclusions

In order to fulfill Bush's MEMEX vision we must seek to offer rapid and flexible access to the contents of a visual, multimedia lifelog. However, as such collections are extremely voluminous and ever-growing, this is particularly challenging. Manual browsing or annotation of the collection to enable retrieval is impractical and we must seek automatic methods to provide reliable annotations to the contents of a visual lifelog. We have documented the process of applying automatic detectors for 27 everyday semantic concepts to a large collection of SenseCam images, and rigorously validated the outcomes. Nine annotators manually judged the accuracy of the output for these 27 concepts on a subset of 95,000 lifelog images spanning five users. We

found that while the concepts' accuracy is varied, depending on the complexity and level of semantics the detector tried to extract from an image, they are largely reliable and offer on average a precision of 57% for positive matches and 93% for negative matches within such a collection.

Furthermore, using the output of the concept detection process, we have been able to identify trends and make inferences into the lifestyles of our 5 users. These inferences were based on the system judgments made for the 27 concepts on an extended collection of almost 220,000 images. By intelligently correlating semantic concepts with previously segmented events or 'activities', we have been able to determine the occurrence of a concept in the users' activities e.g. user 2 has 52 eating events. We have determined through qualitative means that this approach is promising for the identification of concept patterns which occur within an individual's visual lifelog and more generally the concepts of interest and importance for an individual.

These results are particularly encouraging and suggest that automatic concept detection methods translate well to the novel domain of visual lifelogs. Once applied to such a collection it offers the ability to enable a range of opportunities, with the most important being the efficient automatic annotation and retrieval within such a voluminous collection.

Acknowledgements

We are grateful to the AceMedia project and Microsoft Research for support. This work is supported by the Irish Research Council for Science Engineering and Technology, by Science Foundation Ireland under grant 07/CE/I1147 and by the EU IST-CHORUS project. We also would like to extend our thanks to the participants who made their personal lifelog collection available for these experiments, and who partook in the annotation effort.

References

1. Bell, G., Gemmell, J.: A Digital Life. Scientific American, (2007).
2. Bovik, A.C., Clark, M., Geisler, W.S.: Multichannel texture analysis using localized spatial filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(1) (1990) 55–73
3. Bush V.: As We May Think, Atlantic Monthly, 176, 1, pp. 101-108, (July 1945).
4. Byrne, D, Lavelle, B, Doherty, A.R, Jones, G.J.F., Smeaton, A.F. Using Bluetooth and GPS Metadata to Measure Event Similarity in SenseCam Images. In Proc. of IMAI'07, July 2007.
5. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
6. DeVaul R.W., Dunn S.: Real-Time Motion Classification for Wearable Computing Applications, 2001, Project paper, http://www.media.mit.edu/wearables/mithril/realtime.pdf
7. Doherty, A.R., Byrne, D., Smeaton, A.F., Jones, G.J.F., Hughes, M.: Investigating Keyframe Selection Methods in the Novel Domain of Passively Captured Visual Lifelogs. In: Proc. of the ACM CIVR 2008, Niagara Falls, Canada, (2008).
8. Doherty, A.R., Smeaton, A.F.: Automatically Segmenting Lifelog Data Into Events. In Proc. 9th International Workshop on Image Analysis for Multimedia Interactive Services (2008).

9. Fleiss, J. L.: Measuring nominal scale agreement among many raters. Psychological Bulletin, Vol. 76, No. 5 pp. 378—382 (1971)
10. Fuller, M, Kelly, L., Jones G.J.F.: Applying Contextual Memory Cues for Retrieval from Personal Information Archives. Proceedings of PIM 2008 Workshop Florence, Italy, (2008)
11. Geusebroek, J.M.: Compact object descriptors from local colour invariant histograms. In: British Machine Vision Conference, Edinburgh, UK (2006)
12. Geusebroek, J., Smeulders, A.W.M.: A six-stimulus theory for stochastic texture. International Journal of Computer Vision 62(1/2) (2005) 7–16
13. Gurrin, C., Smeaton, A.F., Byrne, D., O'Hare, N., Jones, G.J., O'Connor, N.: An Examination of a Large Visual Lifelog. In AIRS 2008 - Asia Information Retrieval Symposium, pp 537-542, Harbin, China, (2008)
14. Hoang, M., Geusebroek, J., Smeulders, A.W.M.: Color texture measurement and segmentation.Signal Processing 85(2) (2005) 265–275
15. Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K.: SenseCam: A Retrospective Memory Aid. In 8th International Conference on Ubiquitous Computing, pp. 177-193, Orange County, USA (2006)
16. Jiang, Y.G., Ngo, C.W., Yang, J.: Towards Optimal bag-of-features for Object Categorization and Semantic Video Retrieval. In Proceedings of the ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands (2007) 494–501
17. Jurie F., Triggs, B.: Creating efficient codebooks for visual recognition. In IEEE International Conference on Computer Vision, Beijing, China, 2005, pp. 604–610.
18. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A New Method for Graylevel Picture Thresholding using the Entropy of the Histogram. Comp. Vis., Grap., & Image Proc. (1985)
19. Landis, J. R. and Koch, G. G.: The measurement of observer agreement for categorical data. *Biometrics*. Vol. 33, pp. 159-174, (1977)
20. Lee, H., Smeaton, A.F., O'Connor, N.E., Jones, G.F.J.: Adaptive Visual Summary of LifeLog Photos for Personal Information Management, In Proc. 1st Intnl. Workshop on Adaptive Infor. Retrieval, (2006) 22-23.
21. Lin, H.T., Lin, C.J., Weng, R.: A note on Platt's probabilistic outputs for support vector machines. Machine Learning 68(3) (2007) 267–276
22. Naphade, M.R., Kennedy, L., Kender, J.R., Chang, S., Smith, J.R., Over, P., Hauptmann A.: A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. Technical Report RC23612, IBM T.J. Watson Research Center, (2005)
23. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M: The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In ACM Multimedia 2006, pp. 421-430, Santa Barbara, USA, (2006)
24. Snoek, C.G.M., van Gemert, J.C., Gevers, T., Huurnink, B., Koelma, D.C., van Liempt, M.,de Rooij, O., van de Sande, K.E.A., Seinstra, F.J., Smeulders, A.W.M., Thean, A.H.C., Veenman, C.J., Worring, M.: The MediaMill TRECVID 2006 semantic video search engine. In: Proceedings of the 4th TRECVIDWorkshop, Gaithersburg, USA (2006)
25. van Gemert, J.C., Snoek, C.G.M., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Comparing compact codebooks for visual categorization. Computer Vision and Image Understanding (2008) Submitted.
26. Vapnik, V.: The Nature of Statistical Learning Theory. 2nd edn. Springer-Verlag, New York, USA (2000)
27. Wang, D., Liu, X., Luo, L., Li, J., Zhang, B.: Video Diver: generic video indexing with diverse features. In: Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval, Augsburg, Germany (2007) 61–70
28. Yanagawa, A., Chang, S.F., Kennedy, L., Hsu, W.: Columbia university's baseline detectors for 374 LSCOM semantic visual concepts. Technical Report 222-2006-8, Columbia University ADVENT Technical Report (2007)