Multimed Tools Appl (2010) 49:119–144
DOI 10.1007/s11042-009-0403-8

Everyday concept detection in visual lifelogs: validation, relationships and trends

Daragh Byrne · Aiden R. Doherty · Cees G. M. Snoek · Gareth J. F. Jones · Alan F. Smeaton

Published online: 21 October 2009

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 captures on average 3,000 images in a typical day, equating to almost 1 million images per year. It can be 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 domain of visual lifelogs. Our 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 evaluated on a subset of 95,907 images, to determine the accuracy for detection of each semantic concept. We conducted further analysis on the temporal consistency, co-occurance and relationships within the detected concepts to more extensively investigate the robustness of the detectors within this domain.

D. Byrne (✉) · A. R. Doherty · A. F. Smeaton
CLARITY: Centre for Sensor Web Technologies, Dublin City University,
Glasnevin, Dublin 9, Ireland
e-mail: daragh.byrne@computing.dcu.ie

C. G. M. Snoek
Intelligent Systems Lab Amsterdam, University of Amsterdam,
Science Park 107, 1098XG Amsterdam, The Netherlands

G. J. F. Jones
Centre for Digital Video Processing, Dublin City University, Glasnevin, Dublin 9, Ireland

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 notion of digitally capturing our memories is known as lifelogging. While many steps have been taken towards managing such ever-growing lifelogging collections [9, 10, 23], we are still far from achieving on-demand, rapid and easy access. 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, 12]. 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, 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 [26, 31, 35], 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 colour-, texture-, and shape-invariant features [13, 14, 17, 33], combined with supervised machine learning using support vector machines [5, 34]. The emphasis on generic indexing by learning has opened up the possibility of moving to larger concept detector sets [19, 32, 36]. Unfortunately these concept detector sets are optimized for the (broadcast) video domain only, and their applicability to other domains such as visual lifelog collections is unclear but is the focus of this work.

Visual lifelog data, and in particular Microsoft SenseCam data—the source for our investigation—is markedly different from typical video or photographic data and 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 the absence of a lens aperture resulting in many images being much darker or brighter than desired for optimal visual analysis. Also, almost half of the images are generally found to contain non-desirable artefacts such as grain, noise, blurring or light saturation [15]. Thus we have conducted an investigation to determine if semantic concept detection methods translate to the novel domain of lifelogs and to determine the degree of robustness, precision and reliability that can be achieved with these approaches on such collections. This investigation, and the results as reported, build upon work previous reported in Byrne et al. [3]. Further to this prior work, here we present extended results and analysis on the reliability of concept detection within the domain of visual lifelogs. Additionally, we explore some aspects of a lifelog, such as temporal consistency and its spatiotemporal nature that may lead to further enhancements of the robustness of concept detection within lifelog archives.

The rest of this paper is organised as follows: first we outline how we applied concept detection to images captured by the SenseCam lifelogging device (Section 2); then we quantitatively describes how accurate our models are in detecting concepts (Section 3); we next examine the temporal consistency (Section 4) and co-occurences (Section 5); finally we summarise this work and outline potentially interesting future endeavours for concept detection within the domain of lifelogging (Sections 6 and 7).

2 Concept detection requirements in the visual lifelog domain

The major requirements for semantic concept detection on visual lifelogs are as follows: a) the identification of everyday concepts; b) the identification of positive and negative examples; and c) reliable and accurate detection. We now discuss how we followed these steps with respect to lifelog images captured by a SenseCam.

2.1 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 [18]. It is typically hung from a lanyard around the neck and, so it provides a 'first person view' on the activities that the wearer is engaged in (Fig. 1). Anything in the view of the wearer can be captured by the SenseCam owing to its fisheye lens. The SenseCam contains several built-in sensors which are designed to monitor the environment of the wearer. These are: a three-axis accelerometer—to detect movement of the wearer; a passive infrared sensor—to

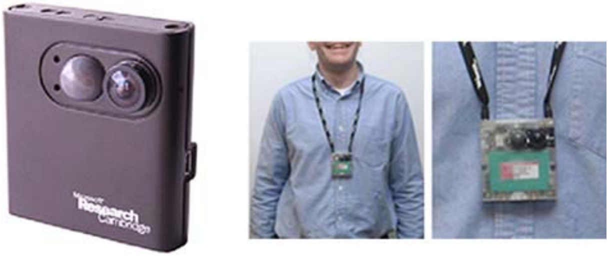

Fig. 1 The Microsoft SenseCam (*Inset*: right as worn by a user)

detect bodies in front of the wearer; light sensor—to detect changes in light level such as when moving from indoors to outdoors; and an ambient temperature sensor. 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, triggers more frequent photo capture. It can capture a typical day without interruption as the battery is sufficient to last for 18 hours and can be recharged fully overnight. 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. Within a year, the lifelog photoset will grow to approximately 1 million images.

Beyond simple triggering of capture by the onboard sensors, the device does not currently support more intelligent or efficient capture of images. As such, the device seeks to capture as much detail about the activities in which a user engages by sampling them at high frequency. With no external control over the decision to capture or the possibility for more selective capture, we must consider means by which we can intelligently determine which of the large number of images produced by this capture mechanism will offer utility. In order to achieve this, we explore a post-processing step, semantic concept detection, through which such understanding of the visual frames can be garnered. We expect that semantic concept detection can ultimately be employed in order to filter, reduce and retrieve the content contained within a visual lifelog. However the motivation of this work is not to immediately offer such functionality but rather to first establish that such techniques translate to this domain and these collections with sufficient success to offer utility.

2.2 Collection overview

In order to evaluate concept detection, we amassed a large and diverse dataset, comprised of 257,518 SenseCam images. These images were gathered by five individual users during five distinct timeframes, and so there was no overlap between the periods captured across 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 physical surroundings. Often collections contained large changes resulting from shifts in location, user behaviour, and/or environments.

2.3 Determining LifeLog concepts

Current approaches to semantic concept detection require the provision of a set of positive and a set of negative labelled exemplar images for each concept. These

Table 1 An overview of the image collection used

User	Total images	Number of positive examples of concepts provided	Days covered
1	79,595	2,180	35
2	76,023	9,436	48
3	40,715	28,023	25
4	42,700	27,223	21
5	18,485	11,408	8
Total	257,518	78,270	137

are then used by a classifier system to train and develop a model for the concept (see Section 2.4. As part of this investigation, we first had to identify the concepts present within the collection for which we wanted to develop detection models, and for which a set of training examples would be collected. In order to determine the typical concepts within the collection, a subset of each user's SenseCam images were visually inspected by playing them sequentially at an accelerated speed. A list of concepts previously used in video retrieval [26, 32] and agreed upon as applicable to a SenseCam collection were used as a starting point. As 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. Next, it was decided that the most representative (i.e. everyday) concepts should be selected and as such the candidates 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. The collection owners were involved throughout the review process and were asked for feedback in negotiating the final selections. 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. We were not motivated to select those concepts which would offer most utility in filtering or retrieval, but rather those which were most likely to occur in all collections and thereby enable robust evaluation of the applicability of semantic concept detection within the domain of visual lifelogs, in which such techniques have not previously been explored. These core concepts are outlined in Fig. 2 using visual examples from the collection. Some concepts are clearly related (e.g. it is logical to expect that 'buildings' and 'outdoors' would co-occur) and as such it is important to note that each image may contain multiple (often semantically related) concepts. This aspect of the collection and of semantic concepts is further discussed in Section 5.

A large-scale manual annotation activity was undertaken to provide the required positive and negative labelled 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 fifth image. As by their nature lifelog images are highly personal, it is important for privacy reasons that it was only the owner of the lifelog images who labels his or her images. Therefore, collection owners annotated their own SenseCam images for the presence of each of the 27 concepts and this provided them an opportunity to remove any portion of their collection they

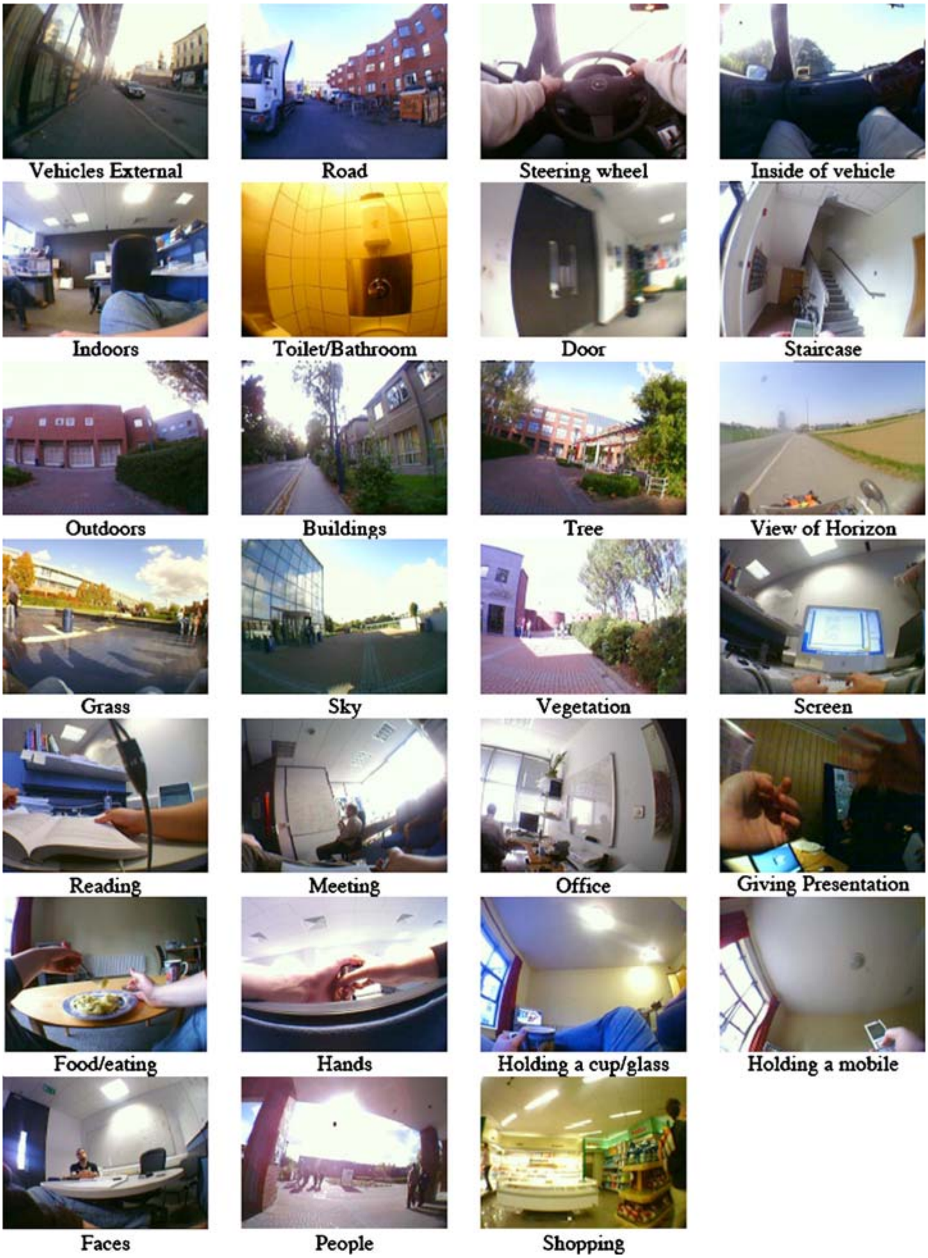

Fig. 2 Visual examples of each of the 27 everyday concepts as detected and validated for the lifelog domain in this paper

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.

Table 2 An outline of the 27 concepts and the number of positive examples per concept and per user

Concept / User	1	2	3	4	5	All
Indoors	1,093	1,439	6,790	6,485	3,480	19,287
Hands	1	17	4,727	3,502	2,402	10,649
Screen (computer/laptop)	7	1,101	4,699	2,628	2,166	10,601
Office	7	78	4,759	2,603	336	7,783
People	0	1,775	573	3,396	889	6,633
Outdoors	250	915	1,248	812	67	3,292
Faces	0	553	101	1,702	662	3,018
Meeting	0	808	0	1,233	355	2,396
Inside of vehicle, not driving (e.g. airplane, car, bus)	257	1,326	420	223	0	2,226
Food (eating)	0	795	349	870	129	2,143
Buildings	140	49	981	621	62	1,853
Sky	0	202	720	525	66	1,513
Road	125	0	231	648	4	1,008
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
Total annotated	16,111	14,787	8,593	8,208	3,697	51,396

2.4 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 stages uses the implementation detailed below.

Supervised Learner: We perceive concept detection in lifelogs as a pattern recognition problem. Given pattern $\mathbf{x}$, 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 [19, 31, 35, 36], we use the Support Vector Machine (SVM) framework [34] for supervised learning of concepts. Here we use the LIBSVM implementation [5] with radial basis function and probabilistic output [24]. 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. [20], which represents an image as a distribution over codewords. We follow [33] 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 colour planes, and therefore well suited for the detection of man-made structures. Both these image features measure coloured texture.

Wiccest features [13] 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 [14] that the complete range of image statistics in natural textures can be well modeled with an integrated Weibull distribution, which in turn can be characterised by just 2 parameters. Thus, 2 Weibull parameter values for the x-edges and y-edges of the three colour 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 colour-texture patches (including e.g. sand, brick, and water), using the accumulated fraction between their Weibull parameters as a similarity measure [33]. 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 $\mathbf{w}$.

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 colour invariance, and 3) construct a histogram. First, the parameters of a Gabor filter consist of orientation, scale and frequency. We use four orientations, $0°, 45°, 90°, 135°$, and two (scale, frequency) pairs: (2.828, 0.720), (1.414, 2.094). Second, colour responses are measured by filtering each colour channel with a Gabor filter. The $\mathcal{W}$ colour invariant is obtained by normalizing each Gabor filtered colour channel by the intensity [17]. Finally, a histogram is constructed for each Gabor filtered colour 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 colour-texture patches, where we use histogram intersection as the similarity measure. Similar to the procedure for $\mathbf{w}$, this yields a codebook feature vector of 120 elements we term $\mathbf{g}$.

Feature and Classifier Fusion: As the visual features $\mathbf{w}$ and $\mathbf{g}$ 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 $\mathbf{w}$ and $\mathbf{g}$ 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 $\mathbf{w}$ and $\mathbf{g}$ without the need to use normalisation or transformation methods. This concatenation yields feature vector $\mathbf{f}$.

For each of the feature vectors in the set $\{\mathbf{w}, \mathbf{g}, \mathbf{f}\}$ 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 labelled examples. To maximize the impact of our labelled 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 [19, 31, 35, 36]. After classifier fusion we obtain our final concept detection score, which we denote $p(\omega_{ij})$.

3 Validation of everyday concept detection

The concept detection process yielded for each frame in the corpus and for each semantic concept, a probability of its presence in the range of 0..1. Although we have trained the detectors using known examples of a given concept, this does not guarantee its reliability, particularly given the variable quality of images in this domain. As such we undertook an evaluation and groundtruthing effort on the system outputs to determine the precision and recall of the developed detectors and additionally to assess the applicability and utility of semantic concept detection within this domain. We now outline the steps taken to manually judge the outputs and then we discuss the outcomes.

3.1 Manual judgements of system outputs

In order to validate the probabilities of a concept's presence, we manually judged a subset of the collection. To facilitate this manual judgement we converted the probabilities in the range of 0..1 into a binary decision of its presence. To make this binary determination, we employed the Kapur automatic thresholding technique [21] which simultaneously selects a threshold value for each concept and divides the collection into those images considered to contain the concept and those which do not. Since this entropy based non-parametric method does not require any training, it can be applied easily to such a broad collection. We consider any images above the threshold value to be positive and those below as negative examples of that concept. Nine participants manually judged a subset of system positive and negative examples for each concept. In order to judge the intercoder reliability—the consistency, and accuracy of each annotator's performance—50 positive and 50 negative examples per concept were randomly selected for judgment by all 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 9×150 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.

An understanding of this 'intercoder agreement' 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 [11]. As such the annotations provided by these participants are consistent and demonstrate high intercoder agreement. Examination at the concept level shows 18 of the 27 concepts had at minimum 0.6 agreement which is substantial according to Landis and Koch [22]. While examination of individual concepts reveals some variability in inter-rater reliability and a 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 large, this may have had the effect 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.

3.2 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 in the 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 concept detector and as such these may be factors in their differing performances. This is outlined in Table 3 where rows are 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 3, 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 in performance 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 Fig. 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. We can explore this point further by examining the bottom 5 performing concepts, namely; 'holdingPhone'; 'holdingCup'; 'meeting'; 'presentation; and 'viewHorizon'. The reason for 'viewHorizon's poor performance is evident from the low number of positive exemplars provided, just three in total. 'HoldingPhone', 'holdingCup' and 'presentation' also perform poorly. At a cursorary level this performance might be attributed to the detector being trained on a relatively low number of visual examples, however, 'toilet' and 'stairs' have lower positive examples yet outperform these concepts. Comparing the sources of the positive examples for these concepts we are provided a cue as to what may be impacting on the detectors' performance. Within the case of 'holdingPhone' and 'holdingCup' we notice that there is a dominance of one user's collection in the provision of positive examples. User 5 provides 63% of the positive examples for the 'holdingPhone' concept, while user 4 provides 84% of the 'holdingCup' examples. 'Toilet' and 'stairs' are less unevenly distributed however they additionally differ

Table 3 Accuracy of detection for each concept (Sorted by 'System Positive Accuracy')

Concept	Examples 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%
Vegetation	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	3,779	29%	99%
viewHorizon	3	3,168	23%	98%

significantly from 'holdingPhone', 'holdingCup' and 'presentation' in their temporal distribution. The positive examples provided for 'holdingPhone', 'holdingCup' and 'presentation' tended to be highly contiguous. A large number of these positive examples represented a sequence of the very visually similar frames, for example, the examples provided for 'presentation' could be aggregated into less than 10 distinct

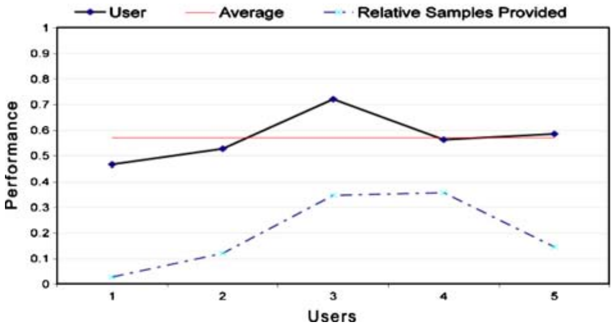

Fig. 3 Performance of all concepts on users' collections

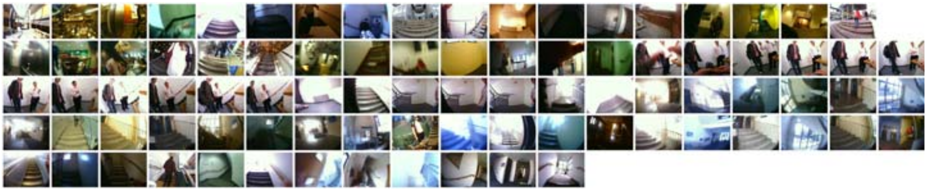

Fig. 4 Visual examples provided for the stairs concept. It can be seen that these examples do not tend to exist in temporally contiguous groups and are visually diverse

and contiguous groups. Conversely, the samples provided for 'stairs' and 'toilet' were more diverse in visual appearance and distribution across the user's collections. This is further illustrated by Figs. 4 and 5. Finally, the 'meeting' concept was emblematic of this issue having a large number of highly contiguous examples, i.e. a large number of examples came from a small number of meetings, and as such this may have yielded a detector more specifically trained to these instances rather than a general case. In summary, we would attribute poor performance of an everyday detector to one or more of the the following issues: a sub-optimal number of positive examples provided for training; a sub-optimal distribution of examples across the user's collection; and/or a sub-optimal diversity in the visual distinctiveness of the provided positive examples (i.e. many highly visually similar examples).

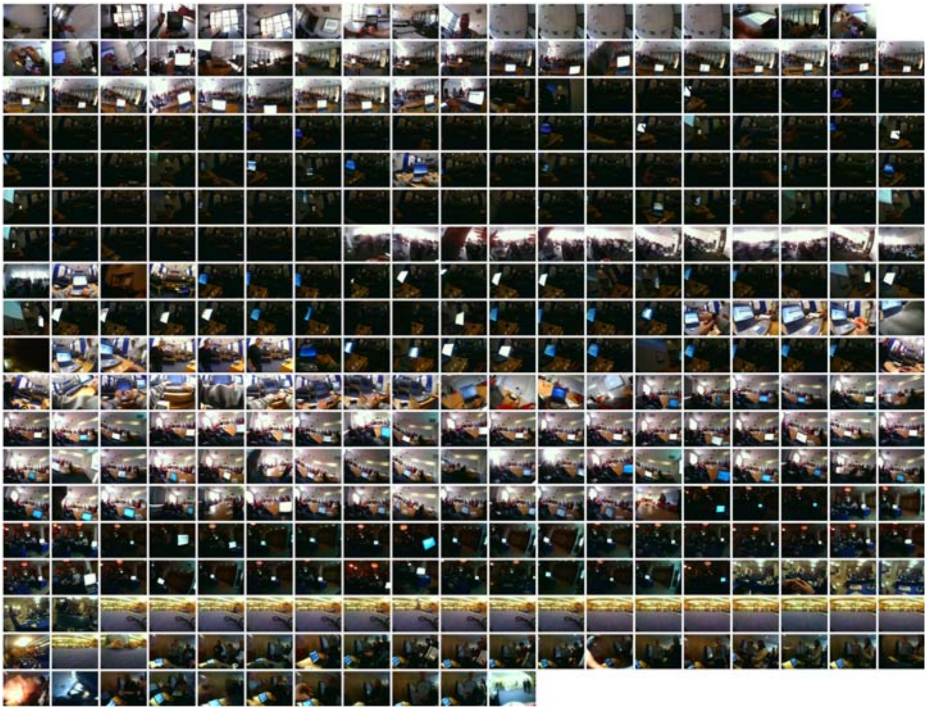

Fig. 5 Visual examples provided for the presentation concept. It can be seen that these examples exist in several temporally contiguous groups and are more homogenous as a result, despite their greater numbers

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 labelled image examples for each concept. We believe the concept detection results on SenseCam images are reliable and the results as presented here demonstrate that the outputs of semantic concept detectors can be safely applied within this domain. We now examine other aspects of semantic concept detection with respect to lifelog collections.

4 Temporal consistency of concepts

In our previous work which explored keyframe selection techniques for SenseCam images, we found that sequences of SenseCam frames can be broadly classified based on their low-level features into one of two groups: visually consistent and visually varied [9]. This property of SenseCam image sequences is illustrated in Fig. 6. Visually consistent frames are present when the user is engaged in one prolonged activity in a fixed location, for example when working at a computer (as illustrated by A), when at a meeting or when watching television. In such instances, we can see that not only do the low-level features remain consistent from frame to frame but the concepts present within each frame are not prone to change either. In sequence A, all of the frames contain the concepts, 'indoor', 'outdoor' and 'screen', while the

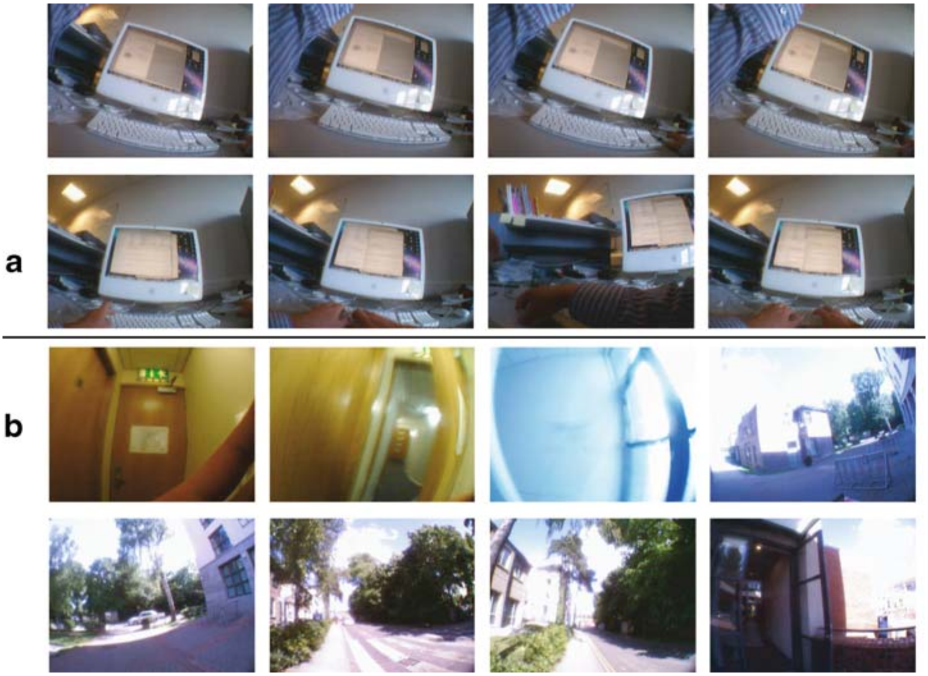

Fig. 6 Visual variance within Sensecam image sequences (**a** depicts a visually consistent sequence, **b** illustrates a visually varied sequence)

final four frames contain ‘hands’ as the user begins typing. Although, the visual features of the images change more significantly from frame-to-frame in visually-varied sequences, concept stability is also present, albeit to a lesser extent. In such sequences, the set of visual concepts present within the frame are much more likely to change in the progression to the next frame. In sequence B, we see a user leave a building and walk across campus. While in the visually consistent sequences the same concepts will be present across the series, concepts here are present only for short bursts e.g. door is present for the first two frames, grass is present for the second and third last frames, etc. Consequently, both visually consistent and visually varied sequences are likely to exhibit some degree of temporal coherence, with respect to the concepts they contain. As the individual concepts within a collection will display different levels of temporal coherence, it is important to quantify the extent to which it is present and by implication the usefulness of this property within such collections.

Temporal coherence is not unique to lifelog collections. Yang and Hauptmann [37] have explored the use of temporal consistency within the TRECVid video collection [29] focusing on broadcast news footage. They define this consistency as *“the tendency that the relevant shots ... appear in temporal proximity”* for a given semantic concept or query. They note that while the degree to which relevant items are temporally proximal is dependent on the topic (or concept), temporal context is nevertheless extremely useful in the prediction of relevance. If a semantic concept or query is known to be highly temporally consistent, then by consulting the prediction on the previous shot, the overall performance of concept detection can be boosted.

In their work, Yang and Hauptman [37] additionally provide three measures to calculate the temporal coherence of a collection for a given concept.

1. *Transitional probability*: is the probability that for a given semantic concept a frame is relevant to the previous frame if its preceding frame is relevant. It is calculated simply as the ratio of the number of consecutive frame pairs to the total number of relevant frames for a given concept. This provides a quantitative measure of the temporal consistency for a given semantic concept.
2. *Marginal probability*: is a comparative measure of the probability of any frame in the collection being relevant for a given semantic concept. It is calculated as the ratio of frames relevant to a concept to the size of the collection. By comparing it with the transitional probability the impact of temporal consistency can be assessed i.e. how much does the transitional probability improve the prediction of a semantics concept’s presence within a given frame.
3. *Pointwise mutual information (PMI)*: It is increasingly probable that relevant frame pairs will occur as the number of relevant frames within the collection increases. The transitional probability is biased towards frequently occurring concepts, and PMI provides a fairer metric of temporal consistency. It is calculated as the log of the ratio of transitional probability against marginal probability.

Using a binary relevance score as calculated previously with Kapur Thresholding (see Section 3) the transitional probability and marginal probabilities for each of the 27 semantic concepts were calculated. We did not remove infrequent concepts as Yang and Hauptmann did given the small set of concepts being used. The distribution of these probabilities was plotted and is presented in Fig. 7. From this we observe

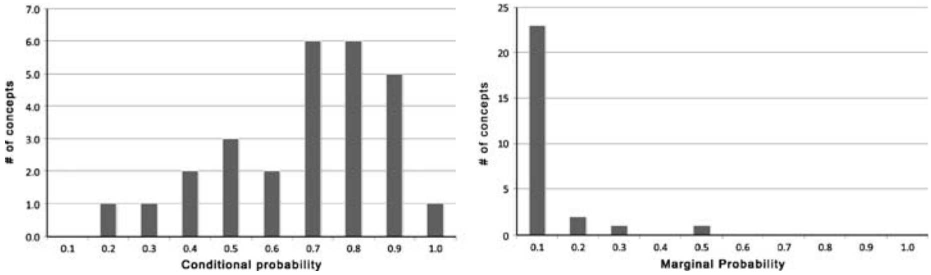

Fig. 7 The distribution of transitional probability (*left*) and marginal probability (*right*) of 27 everyday concepts. Please note the different scales

that the marginal probability for most of the concepts is below 0.1 with an average of 0.07 while the transitional probability demonstrates distribution in much higher ranges, averaging 0.68. For comparative purposes Yang and Hauptman found the TRECVid news video corpus to contain an average marginal probability of 0.038 and a transitional probability averaging at 0.452 for the LSCOM concepts [37]. While this indicates that visual lifelogs may be much more temporally consistent than broadcast video, it should be remembered that Yang and Hauptmann surveyed a far wider range of concepts and additionally filtered to remove infrequently occurring concepts. 194 of 370 LSCOM semantic concepts were used in their evaluation. Additionally, they conducted their evaluation at the shot level while we conduct ours at the image level.

However, for many of the ‘everyday’ concepts, the temporal consistency is extremely high. Six of the concepts show a transitional probability of over 0.85. These include: hands (0.854); indoor (0.903); insideVehicle (0.868); office (0.889); screen (0.886) and steeringWheel (0.966). Only three of the concepts presented a transitional probability poorer than coinflip chance (0.5): holdingPhone (0.398); stairs (0.338); and viewHorizon (0.183). It is interesting to note that these are also the concepts with the lowest number of relevant frames within the corpus, and that they also represent short-duration activities. This is not the case with the maximum values with ‘steeringWheel’ having an average number of relevant frames. This concept had the largest transitional probability and for it almost all relevant frames were transitional pairs. The marginal probability for this concept was 0.027 demonstrating that temporal consistency can aid prediction by almost 36 times. While ‘viewHorizon’ offers the lowest temporal consistency, its marginal probability is almost zero and as such the temporal coherence can also offer it a significant boost in the prediction of relevance.

On average the transitional probability is 94 times larger than the marginal probability. This is reflected in the pointwise mutual information displayed in Table 4. We see that the majority of concepts have a PMI of above 1.5. However, these findings demonstrate that the majority of the 27 everyday concepts display extremely strong temporal consistency. Given the significant difference between the transitional and marginal probabilities, concepts within lifelogs are demonstrated to be highly visually consistent. This is likely to be an inherent trait of this type of collection. There is the possibility to augment and improve the prediction of a concept’s presence within a frame by leveraging temporal consistency.

Table 4 The transitional probability, marginal probability and pointwise mutual information (PMI) for each of the 27 concepts

Concept	Transition probability	Marginal probability	PMI
Buildings	0.758	0.044	1.233
Door	0.469	0.007	1.801
Eating	0.799	0.048	1.218
Face	0.750	0.056	1.124
Grass	0.680	0.012	1.741
Hands	0.854	0.239	0.552
holdingCup	0.596	0.005	2.041
holdingPhone	0.398	0.002	2.252
Indoor	0.903	0.491	0.264
insideVehicle	0.868	0.056	1.192
Meeting	0.780	0.051	1.183
Office	0.889	0.205	0.637
Outdoor	0.818	0.082	0.999
People	0.768	0.138	0.745
Presentation	0.745	0.008	1.990
Reading	0.666	0.016	1.626
Road	0.593	0.014	1.630
Screen	0.886	0.263	0.527
Shopping	0.408	0.004	2.056
Sky	0.753	0.039	1.289
Stairs	0.338	0.001	2.743
steeringWheel	0.966	0.027	1.551
Toilet	0.476	0.002	2.331
Tree	0.719	0.026	1.442
Vegetation	0.696	0.020	1.532
vehiclesExternal	0.507	0.008	1.787
viewHorizon	0.183	0.000	2.896

5 Co-occurance and relationships between concepts

In the previous section we saw that the images in a lifelog are related to those they are temporally aligned with and the concepts they contain exhibit a tendency to be consistent across images. However, within lifelog collections, there is not only a relationship *between* images but also *within* the individual images given that concepts may be semantically related to one another. If we examine the exemplar images for each concept (see Fig. 2) we see that the concepts do not occur in isolation but co-occur within images. For example, it would be expected that an image containing either 'plant', 'tree', 'grass' or 'vegetation' should also contain the 'outdoor' concept; the presence of 'hands' should have a strong relationship with the presence of 'holdingCup' or the 'holdingPhone' concepts; we would also anticipate 'faces' would be found where 'people' occur and vice versa. It seems sensible to assume that the implicit semantic relationships between concepts should be preserved by the concept detection process and that these relationships should then be present in the lifelog collections, providing the detection process is reliable. This, of course, is not a new idea and was proposed by Naphade and Huang [25] among others. Within this section, we explore this as a means by which we can further evaluate the reliability of the detection process.

We examined the occurrences of the 27 concepts within each image in the collection. Again using Kapur thresholding to yield a binary decision for the presence or absence of a given semantic concept within an individual image, we determined the concepts present for each image. Then for each image, we were able to determine concepts which were co-occurrent. Relationships between concepts are bi-directional and this must be considered when calculating the measure of the relationships between them. For example, the relationship of ‘outdoors’ to ‘grass’ is distinct from the relationship of ‘grass’ to ‘outdoors’. To account for this bidirectional relationship, the strength of a relationship from Concept A to Concept B is measured as the ratio of the sum of co-occurences against the total instances of Concept A in the collection. This was calculated for all concept pairs and is presented as a matrix in Table 5.

By examining the co-occurrences and relationships for indoor concepts we can see that they are stable and as expected. We would expect that certain concepts such as ‘office’, ‘presentation’, ‘screen’, ‘stairs’ and ‘toilet’ would have a strong association with ‘indoors’. This is indeed the case with them having a relationship strength to the ‘indoor’ concept of 0.93; 0.83; 0.85; 0.94; and 0.93 respectively. In these cases we can assert that the presence of these concepts in an image is contingent upon the presence of the ‘indoor’ concept. However, the presence of the ‘indoors’ concept does not predicate the appearance of these concepts, with ‘indoors’ having a low-scoring association for them. This is particularly true of stairs for which ‘indoors’ approaches zero association. It is interesting to note the associations between ‘screen’, ‘hands’ and ‘indoor’, each display a high affinity for each other (hands-indoor, 0.94; indoor-hands: 0.45; indoor-screen, 0.45; screen-indoor: 0.85; hands-screen: 0.73; screen-hands: 0.67). This suggests that a large majority of the time spent indoors is in the presence of a screen with hands visible and would indicate that for a large part of the day the users were working with their computer and typing. This demonstrates the potential, as suggested by Naphade and Huang [25], to utilise the relationships displayed between two or more detected concepts to abstract to higher level semantic concepts and infer actions being undertaken within the images.

The co-occurrences are also as expected for outdoor concepts, where again the presence of the ‘outdoor’ concept does not predicate the appearance of ‘grass’, ‘trees’ or ‘vegetation’. However, the appearance of ‘road’ (road-outdoor, 0.97), ‘grass’ (grass-outdoor, 0.88), ‘vegetation’ (vegetation-outdoor, 0.92) or ‘trees’ (trees-outdoor, 0.94) is virtually contingent upon the wearer being outdoors. Additionally, ‘tree’ and ‘vegatation’ also display association to one another (tree-veg, 0.68; veg-tree 0.87). Unlike the other ‘outdoor’ concepts, the presence of sky is conditional upon the presence of ‘outdoor’ (sky-outdoor, 0.93), but ‘outdoors’ also displays a reasonably high affinity to ‘sky’ (outdoor-sky, 0.44) or there is almost a 50-50 chance of ‘sky’ being present where ‘outdoors’ is detected. Other relationships of note include the co-occurences of ‘people’ and ‘faces’ where faces infers that people are present (face-people, 0.93) but the fact that people are present does not necessarily mean faces will be visible (people-faces, 0.38).

The co-occurences also highlight some unexpected relationships within the collections. For example, logically we know that a road will never be found indoors, however, within the corpus ‘road’ was seen have a 0.15 association with ‘indoors’ or put in other words for every 100 images detected to contain road, in 15 of those

Table 5 Matrix of co-occurence for the 27 concepts

	Buildings	Door	Eating	Face	Grass	Hands	holdingCup	holdingPhone	Indoor	insideVehicle	Meeting	Office	Outdoor	People
Buildings		0.03	0.18	0.12	0.17	0.04	–	–	0.03	0.02	–	0.02	0.95	0.16
Door	0.18		0.12	0.15	0.08	0.23	–	0.03	0.97	–	0.08	0.22	0.01	0.17
Eating	0.17	0.02		0.26	0.05	0.42	0.01	–	0.57	0.01	0.01	0.05	0.02	0.36
Face	0.09	0.02	0.22		0.01	0.35	0.06	0.01	0.79	0.17	0.23	0.27	0.05	0.93
Grass	0.61	0.05	0.20	0.04		0.02	–	0.01	0.05	0.01	–	0.04	0.88	0.07
Hands	0.01	0.01	0.08	0.08	–		0.02	0.01	0.94	0.01	0.06	0.57	0.01	0.14
holdingCup	0.01	0.01	0.05	0.63	0.01	0.75		–	0.83	0.07	0.02	0.05	0.03	0.79
holdingPhone	0.08	0.10	0.01	0.23	0.06	0.83	–		0.72	0.12	0.12	0.17	0.14	0.32
Indoor	–	0.01	0.06	0.09	–	0.46	0.01	–		0.01	0.07	0.39	0.01	0.18
insideVehicle	0.02	–	0.01	0.17	–	0.05	0.01	–	0.07		–	0.03	0.02	0.04
Meeting	–	0.01	0.01	0.26	–	0.28	–	0.01	0.67	–		0.37	0.01	0.59
Office	–	0.01	0.01	0.07	–	0.67	–	–	0.93	0.01	0.09		0.01	0.10
Outdoor	0.51	–	0.01	0.03	0.13	0.04	–	–	0.05	0.01	0.01	0.01		0.16
People	0.05	0.01	0.13	0.38	0.01	0.24	0.03	0.01	0.64	0.02	0.22	0.14	0.09	
Presentation	0.03	0.01	0.01	0.21	–	0.06	–	–	0.83	0.03	0.04	0.04	0.02	0.84
Reading	0.01	–	0.01	0.01	–	0.75	–	–	0.85	–	0.08	0.44	0.01	0.08
Road	0.77	–	0.01	0.14	0.23	0.02	–	–	0.15	0.06	–	0.01	0.97	0.11
Screen	–	–	0.01	0.04	–	0.67	–	–	0.85	0.02	0.04	0.62	0.03	0.06
Shopping	0.01	–	0.12	0.08	0.01	0.38	–	–	0.59	0.07	0.01	0.31	0.05	0.24
Sky	0.70	0.02	0.10	0.03	0.27	0.05	–	–	0.03	0.01	–	0.01	0.93	0.18
Stairs	0.05	0.04	0.01	0.01	–	0.06	–	0.01	0.94	0.01	0.01	0.05	0.06	0.08
steerWheel	0.02	–	–	0.14	–	0.02	–	–	0.03	0.16	–	0.02	0.03	0.02
Toilet	0.06	0.35	–	0.17	0.01	0.11	0.01	–	0.93	0.04	0.01	0.16	0.18	0.06
Tree	0.68	0.04	0.14	0.05	0.32	0.05	–	0.01	0.02	0.01	–	0.01	0.94	0.10
Veg	0.67	0.01	0.18	0.05	0.37	0.07	–	0.01	0.04	–	0.02	0.03	0.92	0.14
vehiclesExternal	0.88	–	–	0.01	0.12	0.03	–	–	0.01	–	–	–	0.93	0.09
viewHorizon	0.17	–	–	0.02	0.03	0.08	–	–	0.23	–	0.03	0.08	0.43	0.07

Table 5 (continued)

	Presentation	Reading	Road	Screen	Shopping	Sky	Stairs	steeringWheel	Toilet	Tree	Veg	vehiclesExternal	viewHorizon
Buildings	0.01	–	0.24	0.02	–	0.61	–	0.01	–	0.40	0.31	0.16	–
Door	0.01	0.01	–	0.16	–	0.10	–	–	0.10	0.13	0.03	–	–
Eating	–	–	–	0.08	0.01	0.08	–	–	–	0.08	0.07	–	–
Face	0.03	–	0.03	0.17	0.01	0.02	–	0.07	0.01	0.02	0.02	–	–
Grass	–	–	0.26	0.02	–	0.85	–	–	–	0.68	0.61	0.08	–
Hands	–	0.05	–	0.73	0.01	0.01	–	–	–	0.01	0.01	–	–
holdingCup	–	0.01	–	0.05	–	0.02	–	–	–	0.01	0.01	–	–
holdingPhone	–	–	0.01	0.30	–	0.08	–	–	–	0.08	0.08	0.01	–
Indoor	0.01	0.03	–	0.45	–	–	–	–	–	–	–	–	–
insideVehicle	–	–	0.01	0.08	–	–	–	0.08	–	–	–	–	–
Meeting	0.01	0.02	–	0.22	–	–	–	–	–	–	0.01	–	–
Office	–	0.03	–	0.80	0.01	–	–	–	–	–	–	–	–
Outdoor	–	–	0.16	0.11	–	0.44	–	0.01	–	0.30	0.23	0.09	–
People	0.05	0.01	0.01	0.12	0.01	0.05	–	–	–	0.02	0.02	0.01	–
Presentation		–	0.01	0.51	–	–	–	0.69	–	–	–	–	–
Reading	–		–	0.76	–	0.01	–	0.34	–	–	–	–	–
Road	–	–		0.03	–	0.55	–	0.01	0.02	0.45	0.38	0.16	–
Screen	0.01	0.05	–		–	–	–	–	–	–	–	–	–
Shopping	–	–	–	0.02		0.11	–	–	–	0.01	0.02	–	–
Sky	–	–	0.20	0.01	0.01		–	–	–	0.58	0.45	0.18	–
Stairs	–	–	–	0.03	0.01	0.02		–	–	–	–	–	–
steerWheel	0.19	0.19	–	0.02	–	–	–		–	–	–	–	–
Toilet	0.01	0.02	0.09	0.02	–	0.01	–	0.01		–	–	–	–
Tree	–	–	0.24	0.01	–	0.86	–	–	–		0.68	0.14	–
Veg	–	–	0.26	0.03	–	0.85	–	–	–	0.87		0.10	–
vehiclesExternal	–	–	0.27	0.01	–	0.82	–	–	–	0.43	0.26		–
viewHorizon	–	–	0.03	0.07	–	0.18	–	–	–	0.37	0.07	0.03	

Please note the concept names have been abbreviated to accommodate the large volume of information. The complete names can be referenced in the other tables we present the image was also determined to be ‘indoors’. Also ‘toilet’ was found to have a relationship to ‘road’ in 9 out of every 100 images. However, only a minor number of these logical fallacies exist in the relationships formed by concept co-occurence. While it is likely likely that these incorrect attributions are the result of a thresholding issue, they offer us additional utility by allowing the poorly-performing detectors or thresholds to be identified and corrected.

From the calculated strengths of co-occurences among the concepts we can see that the concept detection process does in fact preserve implicit semantic relationships among the concepts it attributes to an individual image. Most importantly, this finding lends further support to the findings of Section 3 and points to the reliability and robustness of such techniques within the domain of lifelogs. Furthermore, the fact that relationships between the semantic concepts present themselves quite obviously within the collections offers great potential for added value. First, the co-occurance relationships may be used to further enhance the robustness by weighting the probability of a given concept’s occurance based on the occurrences of other concepts within that image. Second, it offers us the potential to abstract and infer higher level semantic concepts based on co-occurences and known relationships, e.g. ‘typing’ from the presence of ‘screen’, ‘hands’ and ‘office’. Finally, and perhaps most usefully, it offers us the ability to automatically extract ontological structures from the collections. The ability to ‘learn’ such structures from the actual occurrences would offer great utility within retrieval applications.

6 Future work

The study reported here 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 are possible.

First the set of concepts presented here represents a very limited set of ‘everyday’ concepts. These are selected as they were generic and expected to be user- and collection-independent. While this is true, the set is highly constrained and does not afford a high degree of utility in practical applications. We must extend this set of concepts to one which is more realistic and covers a more broad range of day-to-day semantic concepts. The number of such concepts needed is an open question. In video retrieval systems, between 100 and 500 concepts are often employed [6, 27, 30]. This upper bound has been established, not as a result of completeness or as effective retrieval is enabled by such numbers, but rather as this is the maximum number of detectors for which annotations are available. For effective retrieval, many more concept detectors may be required. For example, Hauptmann estimates 5,000 detectors would be required [16].

There is also scope to enhance the robustness of concept detection approaches within lifelog archives. As outlined, the images which compose a lifelog collection tend to be temporally consistent in their visual properties and in the concepts they contain. Both prior work [37] and our assertions support the conclusion that this property can be leveraged to further validate the presence of a concept. Likewise and as suggested by Naphade and Huang [25], the expected semantic relations (perhaps formalised within an ontological structure) and the observed semantic relations (as presented in Section 5) offer another means by which the outputs of the semantic detection process can be further enhanced. These relations can be used to upweight the probability of a concept's presence in a image depending on the presence of other concepts or be employed to downweight or remove concepts which are unlikely or non-relevant based on the presence of other concepts.

In addition to the photos the SenseCam captures, it also continually records readings from its onboard sensors (light, temperature and accelerometer sensors.) The measurements taken from these sensors could be useful to augment and enhance the detection of concepts from visual features or to detect wholly new 'activity-centric' concepts as in [7]. Other contextual sources such as nearby Bluetooth devices and GPS location could also be used in augmentation [4]. With a knowledge of location, as in the MediAssist system [28], context-aware concepts may be applied. For example, with a given location, and having detected an image as 'outdoors', weather conditions such as 'windy', 'overcast', or 'raining' may be applied to a image in order to supplement the detected semantic concepts.

Concept-based retrieval has been extremely effective in the domain of digital video [30, 31]. As in video retrieval, these concepts offer the ability to bridge semantic understanding to enable search and location of images relevant to an information need. Retrieval using automatically detected concepts within visual lifelogs should be explored. We plan to undertake evaluations on lifelog collections to assess the utility of such retrieval methods. The performance and utility of concept-based retrieval approaches should also be compared and/or augmented with other methods such as the use of social context [4].

Another area for exploration would be the use of semantic concepts to determine the relative importance of various events. In previous work, the presence of face-to-face conversations and the visual novelty of a given event has been used to automatically determine event importance [8]. We intend to investigate the viability and accuracy of determining event importance through extracted semantic concepts.

Finally, we believe that the exploration of active learning approaches which would combine user-contributed tagging (or folksonomies) with concept detection training, could be undertaken. This would offer a means by which users could create and train new concept detectors as they explore and annotate their collections, allowing efficient and automatic annotation of new content with any available concepts while providing scope and flexibility for a user to personalise their set of concepts.

7 Conclusions

Rapid and flexible access to the contents of a visual lifelog is essential to its utility and usability. However, as such collections are large and ever-growing, this is particularly challenging. Manual browsing or annotation of the collection to enable retrieval is impractical and we should seek automatic methods to provide reliable annotations to the contents of a visual lifelog. We have documented the process of applying automatic detection for 27 everyday semantic concepts to a collection of SenseCam images, and validated the outcomes. Nine annotators manually judged the accuracy of the output for these 27 concepts on a subset of 95,507 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.

Using the output of the concept detection process, we have also explored the temporal consistency, relationships and co-occurences among the detected concepts. On average the transitional probability is 94 times larger than the marginal probability and six of the concepts displayed a transitional probability of over 0.85. Given the magnitude of the transitional probability and the significant difference between it and the marginal probability, we found the 'everyday' semantic concepts within lifelog collections to be highly temporally consistent and coherent. Furthermore, we highlight that concepts may be semantically related to one another. These relationships between the semantic concepts were found to be as expected and this illustrates that the detection process does preserve the implicit semantic relationships among the concepts. Given that the relationships between the semantic concepts present themselves quite obviously within the collections, there is great potential to infer ontological structure, to abstract new or higher-level concepts and to weight the probabilities of occurrence based on these relationships.

These results are particularly encouraging and suggest that automatic concept detection methods translate well to the domain of visual lifelogs.

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 would also 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 (2007) A digital life. Scientific American, New York
2. Bovik A, Clark M, Geisler W (1990) Multichannel texture analysis using localized spatial filters. IEEE Trans Pattern Anal Mach Intell 12(1):55–73
3. Byrne D, Doherty AR, Snoek CG, Jones GG, Smeaton AF (2008) Validating the detection of everyday concepts in visual lifelogs. In: SAMT '08: proceedings of the 3rd international conference on semantic and digital media technologies. Springer, Berlin, pp 15–30
4. Byrne D, Lavelle B, Doherty AR, Jones GJF, Smeaton AF (2007) Using bluetooth and GPS metadata to measure event similarity in sensecam images. In: IMAI'07 - 5th international conference on intelligent multimedia and ambient intelligence, Salt Lake City, pp 1454–1460
5. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm
6. Chang SF, He J, Jiang YG, Khoury EE, Ngo CW, Yanagawa A, Zavesky E (2008) Columbia University/VIREO-CityU/IRIT TRECVid2008 High-Level feature extraction and interactive video search. In: Proceedings of TRECVid workshop, Gaithersburg, 2008
7. DeVaul R (2001) Real-time motion classification for wearable computing applications. Tech. rep., Massachusetts Institute of Technology, MIT, Cambridge
8. Doherty A, Smeaton AF (2008) Combining face detection and novelty to identify important events in a visual lifelog. In: CIT 2008—IEEE international conference on computer and information technology, workshop on image- and video-based pattern analysis and applications, Sydney
9. Doherty AR, Byrne D, Smeaton AF, Jones GJF, Hughes M (2008) Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs. In: CIVR '08: proceedings

of the 2008 international conference on content-based image and video retrieval, Niagara Falls, Canada. ACM, New York, pp 259–268

10. Doherty AR, Smeaton AF (2008) Automatically segmenting lifelog data into events. In: WIAMIS ’08: proceedings of the 2008 ninth international workshop on image analysis for multimedia interactive services, Klagenfurt, Germany. IEEE Computer Society, Washington, DC, pp 20–23
11. Fleiss J (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5): 378–382
12. Fuller M, Kelly L, Jones GJF (2008) Applying contextual memory cues for retrieval from personal information archives. In: PIM 2008 - proceedings of personal information management, workshop at CHI 2008
13. Geusebroek JM (2006) Compact object descriptors from local colour invariant histograms. In: British machine vision conference, vol 3, pp 1029–1038
14. Geusebroek JM, Smeulders AWM (2005) A six-stimulus theory for stochastic texture. Int J Comput Vis 62:7–16
15. Gurrin C, Smeaton AF, Byrne D, O’Hare N, Jones GJF, O’Connor NE (2008) An examination of a large visual lifelog. In: AIRS 2008—Asia information retrieval symposium, Harbin
16. Hauptmann A, Yan R, Lin WH (2007) How many high-level concepts will fill the semantic gap in news video retrieval? In: CIVR ’07: proceedings of the 6th ACM international conference on image and video retrieval. ACM, New York, pp 627–634
17. Hoang MA, Geusebroek JM, Smeulders AWM (2005) Color texture measurement and segmentation. Signal Process 85(2):265–275
18. Hodges S, Williams L, Berry E, Izadi S, Srinivasan J, Butler A, Smyth G, Kapur N, Wood K (2006) SenseCam: a retrospective memory aid. In: UbiComp - 8th international conference on ubiquitous computing, Calif., USA
19. Jiang YG, Ngo CW, Yang J (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval. In: CIVR ’07: proceedings of the 6th ACM international conference on image and video retrieval. ACM, New York, NY, USA, pp 494–501
20. Jurie F, Triggs B (2005) Creating efficient codebooks for visual recognition. In: Computer vision, 2005. ICCV 2005. Tenth IEEE international conference on 1, 604–610, vol 1
21. Kapur J, Sahoo P, Wong A (1985) A new method for gray-level picture thresholding using the entropy of the histogram. Comput Vis Graph Image Process 29(3):273–285
22. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
23. Lee H, Smeaton AF, O’Connor NE, Jones GJ (2006) Adaptive visual summary of lifelog photos for personal information management. In: AIR Workshop—1st international workshop on adaptive information retrieval, Glasgow, pp 22–23
24. Lin HT, Lin CJ, Weng R (2007) A note on Platt’s probabilistic outputs for support vector machines. Mach Learn 68(3):267–276
25. Naphade H, Huang T (2001) A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans Multimedia 3(1):141–151
26. Naphade MR, Kennedy L, Kender JR, Chang SF, Smith JR, Over P, Hauptmann A (2005) A light scale concept ontology for multimedia understanding for TRECVid 2005. Tech. rep., In IBM Research Technical Report
27. Natsev A, Jiangy W, Merlery M, Smith JR, Tesic J, Xie L, Yan R (2008) IBM research TRECVid-2008 video retrieval system. In: Proceedings of TRECVid workshop, 2008, Gaithersburg
28. O’Hare N, Lee H, Cooray S, Gurrin C, Jones GJF, Malobabic J, O’Connor NE, Smeaton AF, Uscilowski B (2006) MediAssist: using content-based analysis and context to manage personal photo collections. In: CIVR2006 - 5th international conference on image and video retrieval. Springer, Tempe, pp 529–532
29. Smeaton A, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval. ACM, New York, pp 321–330
30. Snoek CGM, Everts I, van Gemert JC, Geusebroek JM, Huurnink B, Koelma DC, van Liempt M, de Rooij O, van de Sande KEA, Smeulders AWM, Uijlings JRR, Worring M (2007) The MediaMill TRECVid 2007 semantic video search engine. In: Proceedings of TRECVid workshop, Gaithersburg, 2007

31. Snoek CGM, van Gemert JC, Gevers T, Huurnink B, Koelma DC, van Liempt M, de Rooij O, van de Sande KEA, Seinstra FJ, Smeulders AWM, Thean AHC, Veenman CJ, Worring M (2006) The MediaMill TRECVID 2006 semantic video search engine. In: Proceedings of the TRECVID workshop, Gaithersburg
32. Snoek CGM, Worring M, van Gemert JC, Geusebroek JM, Smeulders AWM (2006) The challenge problem for automated detection of 101 semantic concepts in multimedia. In: MULTIMEDIA '06: proceedings of the 14th annual ACM international conference on multimedia, Santa Barbara, CA, USA. ACM, New York, pp 421–430
33. van Gemert JC, Snoek CGM, Veenman CJ, Smeulders AWM, Geusebroek JM (2009) Comparing compact codebooks for visual categorization. Comput Vis Image Underst. doi:10.1016/j.cviu.2009.08.004
34. Vapnik VN (2000) The nature of statistical learning theory, 2nd edn. Springer, New York
35. Wang D, Liu X, Luo L, Li J, Zhang B (2007) Video diver: generic video indexing with diverse features. In: MIR '07: proceedings of the 9th ACM international workshop on workshop on multimedia information retrieval, Augsburg, Germany. ACM, New York, pp 61–70
36. Yanagawa A, Chang SF, Kennedy L, Hsu W (2007) Columbia University's baseline detectors for 374 LSCOM semantic visual concepts. Tech. rep., Columbia University
37. Yang J, Hauptmann AG (2006) Exploring temporal consistency for video analysis and retrieval. In: MIR '06: proceedings of the 8th ACM international workshop on multimedia information retrieval, Santa Barbara, pp 33–42

Daragh Byrne is a PhD Student with the Centre for Digital Video Processing (CDVP) at Dublin City University (DCU), under the supervision of Dr. Gareth Jones. He is funded by a research scholarship awarded by the Irish Research Council for Science, Engineering and Technology (IRCSET) since 2007 and his work is also supported by the CLARITY centre. He holds a M.Res. degree in Design and Evaluation of Advanced interactive Systems from Lancaster University and a BSc. in Computer Applications from DCU. Since joining the CDVP, he has published over 25 scientific papers and has been actively involved in the lifelogging domain. The majority of his work focuses in this area and his PhD thesis will focus on creating digital narratives from multimodal lifelog content.

Aiden R. Doherty is a postdoctoral researcher in CLARITY: Centre for Sensor Web Technologies. His research interests cover lifelogging, sports & personal health IT applications, location aware computing, multimedia information management, brain wave readers, and data mining. He has approximately 15 publications, has given 5 seminar talks and had a three month internship with Microsoft Research in Redmond, WA, USA. He has also distributed a lifelogging browsing system to over 10 universities to enable the field proceed more efficiently. He holds a 1:1 BSc in computer science from the University of Ulster, and at the age of 24 a PhD from Dublin City University where he was a government of Ireland research scholar.

Cees G. M. Snoek received the M.Sc. degree in business information systems (2000) and the Ph.D. degree in computer science (2005) both from the University of Amsterdam, The Netherlands. He is currently a senior researcher at the Intelligent Systems Lab Amsterdam, funded by a young talent (VENI) grant from the Netherlands Organization for Scientific Research. He was a Visiting Scientist at Informedia, Carnegie Mellon University, USA in 2003. His research interests focus on video retrieval. He has published over 70 refereed scientific papers. Dr. Snoek is a lead researcher of the award-winning MediaMill Semantic Video Search Engine, which is a consistent top performer in the yearly NIST TRECVID evaluations. He is initiator and co-organizer of the annual VideOlympics, and a lecturer of post-doctoral courses given at international conferences and European summer schools. Dr. Snoek is a member of ACM and IEEE.

Gareth J. F. Jones is a Senior Lecturer in the School of Computing and a Principal Investigator in the Centre for Digital Video Processing at Dublin City University. His research interests include multimedia and multilingual information access and retrieval, mobile and personal computing technologies, information visualisation and affective computing. He has published over 180 scientific papers describing this work. He has previously held posts at University of Exeter and University of Cambridge, U.K. In 1997, he was a Toshiba Fellow at the Toshiba Corporation Research and Development Center in Kawasaki, Japan, and in 2002 a Visiting Scientist to the Informedia project at Carnegie Mellon University, U.S.A. and a JSPS Visiting Fellow at the National Institute of Informatics, Tokyo, Japan. He holds B.Eng and PhD degrees in Electrical and Electronic Engineering from the University of Bristol. He is a member of the ACM, UK IET and IEEE Computer Society.

Alan F. Smeaton is a Professor of Computing at Dublin City University where he is the Deputy Director of CLARITY: Centre for Sensor Web Technologies. His research interests cover indexing and content-based retrieval of information in all media, text, image, audio and especially digital video and now the focus of his work is in information access for human digital memory applications. His major research funding is in the area of information analysis and access, particularly for digital video, and has also received funding for research in digital libraries, music IR and in web searching. He is a founder and coordinator of TRECVid, an annual benchmarking activity for content-based video operations which has operated annually since 2001 and involves almost 80 research groups worldwide. Alan Smeaton is a member of the ACM and IEEE Computer Society, and a Fellow of the Irish Computer Society.