Characterizing Everyday Activities from Visual Lifelogs based on Enhancing Concept Representation

Peng Wang[1*], Lifeng Sun[1], Shiqiang Yang[1], Alan F. Smeaton[2], Cathal Gurrin[2]

[1] *National Laboratory for Information Science and Technology*
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China

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

Abstract

The proliferation of wearable visual recording devices such as SenseCam, Google Glass, etc. is creating opportunities for automatic analysis and usage of digitally-recorded everyday behaviour, known as visual lifelogs. Such information can be recorded in order to identify human activities and build applications that support assistive living and enhance the human experience. Although the automatic detection of semantic concepts from images within a single, narrow, domain has now reached a usable performance level, in visual lifelogging a wide range of everyday concepts are captured by the imagery which vary enormously from one subject to another. This challenges the performance of automatic concept detection and the identification of human activities because visual lifelogs will have such variety of semantic concepts across individual subjects. In this paper, we characterize the everyday activities and behaviour of subjects by applying a hidden conditional random field (HCRF) algorithm on an enhanced representation of semantic concepts appearing in visual lifelogs. This is carried out by first extracting latent features of concept occurrences based on weighted non-negative tensor factorization

*Corresponding author at: Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China. Tel: +86 -10 -62786910

Email addresses: pwang@tsinghua.edu.cn (Peng Wang[1]), sunlf@tsinghua.edu.cn (Lifeng Sun[1]), yangshq@tsinghua.edu.cn (Shiqiang Yang[1]), alan.smeaton@dcu.ie (Alan F. Smeaton[2]), cathal.gurrin@computing.dcu.ie (Cathal Gurrin[2])

(WNTF) to exploit temporal patterns of concept occurrence. These results are then input to an HCRF-based model to provide an automatic annotation of activity sequences from a visual lifelog. Results for this are demonstrated in experiments to show the efficacy of our algorithm in improving the accuracy of characterizing everyday activities from individual lifelogs. The overall contribution is a demonstration that using images taken by wearable cameras we can capture and characterize everyday behaviour with a level of accuracy that allows useful applications which measure, or change that behaviour, to be developed.

Keywords: Lifelogging, Assistive living, SenseCam, Activity classification, Wearable camera

---

1. Introduction

There is growing interest in creating large volumes of personal, first-person video or long duration image sequences, for lifelogging or quantified-self types of applications. These use wearable visual recording devices like Google Glass or Microsoft's SenseCam. *Visual lifelogging* is the term used to describe a class of personal sensing and digital recording of all of our everyday behaviour which employs wearable cameras to capture image or video sequences of everyday activities. As the enabler for visual lifelogging, camera-enabled sensors are used in wearable devices to record still images [41] or video [14, 31, 4] taken from a first-person view, i.e. representing the subject's view of everyday activities. Visual lifelogging has already been widely applied in assistive living applications including aiding human memory recall, diet monitoring, chronic disease diagnosis, recording activities of daily living and so on. Example visual lifelogging projects include Steve Mann's WearCam [30, 31], the DietSense project at UCLA [39], the WayMarkr project at New York University [5], the InSense system at MIT [4], and the IMMED system [32]. Microsoft Research catalysed research in this area with the development of the SenseCam [13, 41] which was made available to other research groups in the late 2000's.

In terms of sensing devices, visual lifelogging can be categorized roughly into in-situ lifelogging and wearable lifelogging. In-situ lifelogging can be described simply as lifelogging in instrumented environments such as homes or workplaces. This means that human activities can be captured through video sensors installed in the local infrastructure [34]. Typical use of video

sensors for in-situ lifelogging also includes work as reported in [18, 16, 1, 49, 17] and [2]. [18] proposed a depth video-based activity recognition system for smart spaces based on feature transformation and HMM recognition. Similar technologies are applied in other work by the same authors in [16] and [1] which can recognise human activities from body depth silhouettes. In related work by [49], depth data is utilised to represent the external surface of the human body. By proposing the body surface context features, human action recognition is robust to translations and rotations. As with Jalal's work in [17, 2], Song's work [49] still depends on static scenes with an embedded sensing infrastructure. Current activity recognition in such settings usually assume there is only one actor in the scene and how these solutions can scale up to more realistic and challenging settings such as outdoors are difficult.

To alleviate such challenges, we focus on activity recognition within non-instrumented environments using wearable visual sensing. In wearable lifelogging, the sensing devices are portable and worn directly by the subjects and can include head-mounted cameras in work by [14] and [31] or cameras mounted on the front of chests in work by [4] and by [41]. In [15], the key issues and main challenges in generating wearable diaries and lifelogging systems are discussed. In [12], other sensors such as accelerometers, GPS, image and audio are recorded using a smartphone and applied in an application based on annotating daily activities. Though effective to a limited extent, a direct mapping from low-level features like colours and textures to semantic labels lacks flexibility in characterizing the semantics of activities such as understanding occurrences of scenes, objects, etc. in images. Recent work in [48] has also highlighted the same problem.

As a new form of multimedia, the effective management of large visual lifelogs requires semantic indexing and retrieval, for which we can use the preliminary work already done in other domains. State-of-the-art techniques for image/video analysis use statistical approaches to map low-level image features like shapes and colours to high-level semantic concepts like "indoor", "dog" or "walk". According to the TRECVid benchmark [44], acceptable detection results have been achieved, particularly for concepts for which there exists enough annotated training data. Introducing automatic detection of semantic concepts from visual lifelogs enables searching through those lifelogs based on their content and this is particularly useful in characterizing everyday life patterns [6, 9]. However, because of the wide variety of activities that people usually engage in and the differences in those activities from person to person, a very wide range of semantic concepts can appear in vi-

sual lifelogs, which increases the challenges in developing automatic concept detectors from which we can detect everyday activities. Moreover, due to subjects' movements as lifelog images are captured, even images captured passively within the same lifelogged event may have significant visual differences. This poses burdens on the characterization of activities based on the detected concepts, especially in applications where the detection of everyday behaviour is to be done in near real-time.

Figure 1: The Microsoft SenseCam as worn by subjects.

The SenseCam, shown in Fig. 1, is a sensor-augmented wearable camera designed to capture a digital record of the wearer's day by recording a series of images and a log of sensor data. It captures the view of the wearer from a fisheye lens and pictures are taken at the rate of about one every 50 seconds without the trigger of other sensors. The on-board sensors for measuring ambient light levels, movement, and the presence of other people through a passive infra-red sensor, are also used to trigger additional capture of pictures when sudden changes are detected in the environment of the wearer as well as to prevent images being captured when the wearer, and the SenseCam, are being moved which would result in blurring of images. SenseCam has been shown to be effective in supporting recall of memory from the past for individuals [41, 43], as well as having applications in diet monitoring [36], activity detection [56], sports training [35], etc. Due to its advantages of multiple sensing capabilities, light weight and unobtrusive logging with a

long battery life, we employ SenseCam as a wearable device to log details of subjects' everyday lives.

Temporal patterns of concept occurrence can characterize image sequences, but at a higher level. Consider the "cooking" activity, where visual concepts like "fridge", "microwave", "oven" often occur in sequence and frequently interact with the concept of "hands". For example, "opening fridge" is typically observed before "starting microwave". Such patterns can also be regarded as temporal semantics of concepts. To deal with such concept temporal semantics, the major contributions of this paper can be highlighted as: first, we proposed a time-aware concept detection enhancement algorithm based on weighted non-negative tensor factorization (WNTF) for which a multiplicative solution is derived. The effectiveness of this factorization method is also proven. The second contribution is an everyday activity characterization based on hidden conditional random fields (HCRF), proposed by merging time-varying dynamics of concept attributes.

The rest of the paper is organized as follows: in Section 2 we present related work on concept detection and event processing used to drive a characterization of everyday activities. An overview of our proposed solution is presented in Section 3. In Section 4, we describe tensor factorization approaches to tackle the concept enhancement problem at a frame-level of concept indexing. This is followed by an HCRF-based algorithm to combine concept semantics from a frame level for higher-level activity characterization in Section 5. The experimental implementation and analysis of our results are presented in Section 6. Finally, we close the paper with conclusions and pointers to future work.

2. Related Work

2.1. Automatic Concept Detection and Enhancement

Compared to low-level features like colour, texture, shape, etc. which in their raw form do not convey much meaning, semantic concepts can more usefully express the real content of any visual media as high-level features such as "indoor", "outdoor", "vegetation", "computer screen", etc. Semantic concepts can be automatically detected, providing a meaningful link between low-level image features and user interpretation of the media. In assistive technology applications, the appearance of semantic concepts can reveal underlying human behaviour, differences in behaviour from person to person, or between an individual person and what we would call normal behaviour,

and if a lifelog is generated for a long enough period, patterns of appearance of semantic concepts can reveal gradual shifts in behaviour such as those associated with degenerative human conditions like Parkinson's disease of forms of dementia.

The state-of-the-art approach to detecting semantic concepts is to apply a suite of discriminative machine learning algorithms such as Support Vector Machines (SVMs) to decide on the presence or absence of each concept given the extracted low-level features [47]. SVMs have been demonstrated to be an efficient framework by many research groups in concept detection [29, 7, 46] and are particularly suited to highly imbalanced classes in terms of the uneven distribution of semantic concepts in visual media.

In many concept detection approaches, applications of machine learning assume that classifiers for a set of concepts are independent of each other, and equally weighted in terms of importance. Yet, intrinsic relationships among concepts are neglected under this assumption so we end up with multiple isolated binary classifiers thus not taking advantage of inter-concept semantics [25]. This is likely to result in misclassification or inconsistencies. Since the accuracy of a concept detector/classifier is an important factor in the provision of satisfactory solutions to indexing visual media, it is widely accepted that detection accuracy can be improved if concept correlation can be exploited in some way. Multi-label training methods such as that proposed in [37] try to learn all concepts from one integrated model, but the direct shortcoming is the lack of flexibility, which means the learning stage needs to be repeated when the concept lexicon is changed or in our case, when the visual lifelog for a new subject is to be analyzed. In other related work, researchers have used ensembles of classifiers like the multi-SVM (MSVM) approach which efficiently handles the issue of an imbalance in the relative sizes of positive and negative classes (concepts) when dealing with concepts in visual media [40]. The idea here is to divide the data into balanced subsets of positive and negative training data, train an SVM on each subset and apply a fusion function to the output of the individual SVMs. While this *divide and conquer* approach is appealing, it greatly increases the processing time required, though recent developments in machine learning such as kernel optimization [26] or approximate learning [22] might alleviate this.

Because concept detection scores obtained by specific binary detectors allow independent and possibly specialized classification techniques to be leveraged for each concept [45], detection enhancement using post-processing also attracts research interest based on exploiting concept correlations inferred

from pre-constructed knowledge [58, 21] or annotation sets [51, 52, 20, 19]. These methods depend highly on external knowledge such as WordNet or other training data. When concepts do not exist in a lexicon or extra annotation sets are insufficient for correlation learning because of the limited size of a corpus or the sparsity of annotations, these methods cannot adapt to such situations. In a state-of-the-art refinement method for the indexing of TV news video by [20, 19], the authors combined the training procedure and the knowledge inference together by learning the concept graph from the training set.

While the work outlined above is making progress in the task of multi-concept indexing of visual media, none of the methods existing to date are able to use any of the temporal semantics that might be part of the collection of visual information. In the domain of visual lifelogging this is an importance feature in trying to determine human activities where unlike other domains, the lifelog collection has a strict linear temporal dimension as it represents a recording of the continuous lifelog of a single individual or subject.

2.2. Concept-Driven Activity Characterization

The purpose of visual lifelogging in our work is that it is an application where we try to characterize activities or events for a human subject in his/her everyday life. In other applications of lifelogging such as memory aids, work-related recording, and so on, a full understanding of activities is usually necessary yet in work in those areas there is still little information representing the semantics of human activities. Discovering activities where there is such little metadata represents a real challenge, especially where we are dealing with long-term lifelog data.

In visual lifelogging, much work has been done on activity processing such as automatic event segmentation [10], event representation [53], life pattern analysis [23], event enhancement [11] and so on. To drive these applications, all of the work uses images from wearable cameras and some additionally uses other captured metadata such as location from GPS, date and time. Recent work on semantic learning from lifelog data has also shown promising results in activity-related concept detection [6] where semantic indexing has shown potential for relating low-level visual features to high-level semantic concepts (such as indoors, outdoors, people, buildings, etc.) using supervised machine learning techniques. This is then applied in [9] to learn lifestyle traits from lifelogs collected by different users, based on automatically detected everyday concepts.

Although the effectiveness for many of the above approaches is satisfactory for some tasks like diet monitoring, we still lack accurate indexing and retrieval methods to localize activities of interest from a large volume of lifelog data. To address such a challenge, concept-based event detection has attracted much attention from researchers. In work by [50], a rule-based method is proposed to generate textual descriptions of video content according to concept classification results. The authors also found that although state-of-the-art concept detections are far from perfect, they are still able to provide useful clues for event classification. Work in [33] employed an intermediate representation of semantic model vectors trained from SVMs, as a basis for detecting complex events, and revealed that this representation outperforms, and is complementary to, other low-level visual descriptors for event modeling. Similar work has also been carried out in lifelogging using concept detections to characterize everyday activities as reported by the authors of this paper [56]. Activity recognition as presented in [56] is also built on the basis of underlying detection of semantic concepts.

3. Overview of Solution Framework

Our approach to characterizing everyday activities for personal applications is based on detection of semantic concepts from a series of images taken from events which have been automatically segmented based on the technique introduced in [28]. An event corresponds to a single everyday activity in the subject's day such as watching TV, commuting to work, or eating a meal, with an average stream of between 20 and 40 such events of varying duration, in a typical day.

Fig. 2 shows the paradigm of utilising concept temporal dynamics for high-level activity detection, in which typical indoor activities like "cooking", "watching TV", etc. are demonstrated, as well as the corresponding trajectories. The concept detection results temporally aligned with these activities are depicted as "$\surd$" and "$\times$" in the diagram, to represent presence and absence of concepts, respectively. It is important to note that the concept detection shown in Fig. 2 does have errors and this can affect further analysis to various degrees. Therefore, in order not to propagate these errors into the subsequent analysis for activity and behaviour characterization, the original concept detections are revised and enhanced in a time-aware manner based on WNTF as presented later in Section 4.

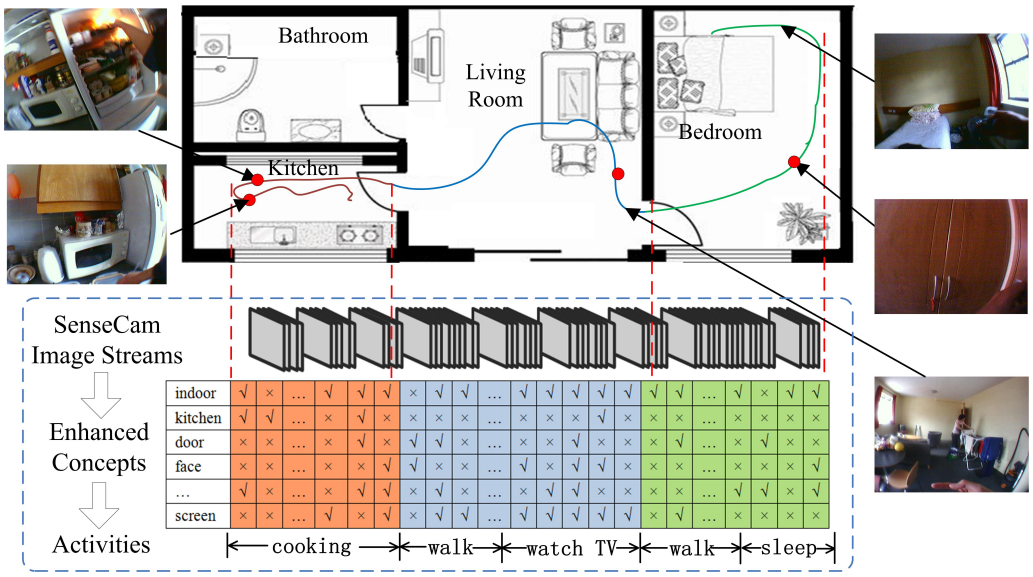


Figure 2: Framework paradigm for proposed solution.

Images taken from a single subject's lifelog are tightly related to those other images with which they are aligned temporally, i.e. they are related to images from around the same event. When a user remains in the same location and is engaged in a prolonged activity such as watching TV in their living room or preparing a meal in their kitchen, the contents of successive captured images are very similar, visually. In this case, the temporal consistency of certain types of concepts like "indoor", "screen" and "hands" can be viewed as cues for concepts like "using computer" while "pages" and "hands" suggests a "reading" activity. Even though some activities require the user to be changing their location all the time like "walking" and "doing housework", the dynamics of concepts present in the activity will still show some patterns, such as the frequent appearance of "road" for the "walking" activity in an urban environment, or the transitions between "kitchen" and "bathroom" for the "housework" activity. Under this assumption, the concepts contained in these images can reflect significant temporal patterns, characterizing the semantics of the activities they represent. An activity characterization algorithm based on HCRF is employed in our work to model the dynamic concept occurrence patterns described above, which is introduced later in Section 5.

4. Time-Aware Concept Detection Enhancement

Co-occurrence and re-occurrence patterns for concepts are a reflection of the contextual semantics of concepts since everyday concepts usually co-occur within images rather than in isolation. In some potentially long-duration activities like "using computer", "driving", etc., the indicating concepts may appear frequently and repeatedly in the first-person view. In this section, the modeling of contextual semantics is elaborated based on time-aware tensor decomposition.

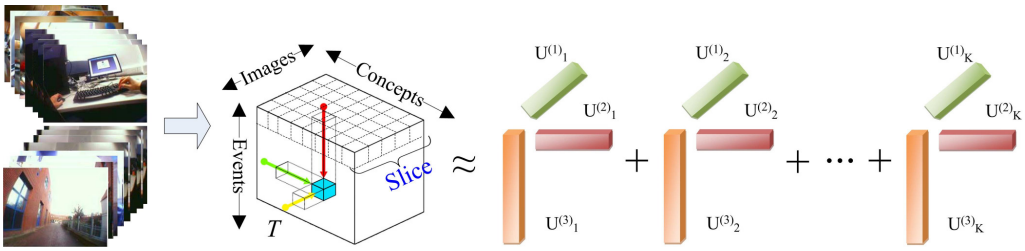


Figure 3: NTF-based concept detection enhancement framework.

To avoid information loss from event segmentation and utilise the temporal features reflected in different events, a tensor is employed to formalize the problem given its merit in representing the structure of multidimensional data more naturally. The procedure for concept tensor construction and factorization is shown in Fig. 3. As illustrated, our approach treats concept detection results in a way which has the advantage of preserving local temporal constraints using a series of two-dimensional slices. Each slice is a segmented part of an event and is represented by a confidence matrix. The slices are then stacked one below another to construct a three-dimensional tensor which preserves the two-dimensional characters of each segment while keeping temporal features along the event dimension and avoids significant loss of contextual information.

4.1. WNTF-based Detection Enhancement

Assume each two-dimensional slice is a segment of N visual lifelog images, each of which is represented by a vector of M concept detection confidences (i.e. concept vectors $(c_{ij})_{M\times 1}, 1 \leq j \leq M$ for the i-th image). The constructed concept detection tensor T has the dimensionality of $N \times M \times L$

for events with L time intervals and N neighborhood lifelog images in each slice. The task of WNTF is to discover the latent features to represent the three components of confidence tensor T. For this purpose, we approximate tensor T as a sum of 3-fold outer-products with rank-K decomposition $\hat{T} = \sum_{f=1}^{K} U_{\cdot f}^{(1)} \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)}$, which means that each element $\hat{T}_{ijk} = \sum_{f=1}^{K} U_{if}^{(1)} U_{jf}^{(2)} U_{kf}^{(3)}$.

This factorization can be solved by optimizing the cost function defined to qualify the quality of the approximation. Similar to work described in [57], we also employ the weighted cost function to distinguish the contribution of different concept detectors, which can be formalized as

$$\begin{aligned} F &= \frac{1}{2}\|T - \hat{T}\|_W^2 = \frac{1}{2}\|\sqrt{W} \circ (T - \hat{T})\|_F^2 \\ &= \frac{1}{2}\sum_{ijk} W_{ijk}(T_{ijk} - \sum_{f=1}^{K} U_{if}^{(1)} U_{jf}^{(2)} U_{kf}^{(3)})^2 \\ &\text{s.t. } U^{(1)}, U^{(2)}, U^{(3)} \geq 0 \end{aligned} \tag{1}$$

where $\circ$ denotes element-wise multiplication, $W = (W_{ijk})_{N \times M \times L}$ denotes the weight tensor and $\|\cdot\|_F^2$ denotes the Frobenius norm, i.e., the sum of squares of all entries in the tensor. To obtain a reconstruction of the underlying semantic structure, the weights in Eqn. (1) need to be set in terms of concept accuracy. Because each confidence value T_{ijk} in T denotes the probability of concept c_j occurring in the image, estimating the existence of c_j is more likely to be correct when T_{ijk} is high enough, which is also adopted in [24] and in [57] under the same assumption that the initial detectors are reasonably reliable if the returned confidences are larger than a predefined value *threshold*. After factorization, the refinement can be expressed as a fusion of the two confidence tensors:

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

A gradient descent method can be applied for optimizing the solution to this problem, implemented by updating each matrix $U^{(t)}$ in the opposite direction to the gradient at each iteration through

$$U^{(t)} \leftarrow U^{(t)} - \alpha_{U^{(t)}} \circ \partial F / \partial U^{(t)}, t = 1, 2, 3 \tag{3}$$

According to [57], the cost function differential with respect to an element $U_{if}^{(1)}$ can be represented as

$$\partial F/\partial U_{if}^{(1)} = \sum_{jk}(W \circ \hat{T})_{ijk}U_{jf}^{(2)}U_{kf}^{(3)} - \sum_{jk}(W \circ T)_{ijk}U_{jf}^{(2)}U_{kf}^{(3)} \quad (4)$$

By employing $\alpha_{U^{(1)}}$ as the form

$$\alpha_{U_{if}^{(1)}} = U_{if}^{(1)}/\sum_{jk}(W \circ \hat{T})_{ijk}U_{jf}^{(2)}U_{kf}^{(3)} \quad (5)$$

where / denotes element-wise division, and substituting into Eqn. (3), we obtain the multiplicative updating rule [27, 42] as

$$U_{if}^{(1)} \leftarrow U_{if}^{(1)}(\sum_{jk}(W \circ T)_{ijk}U_{jf}^{(2)}U_{kf}^{(3)})/(\sum_{jk}(W \circ \hat{T})_{ijk}U_{jf}^{(2)}U_{kf}^{(3)}) \quad (6)$$

The updating of $U^{(2)}$ and $U^{(3)}$ can be achieved in a similar manner. Note that it is not difficult to prove that under such updating rules, the cost function in Eqn. (1) is non-increasing in each optimization step.

4.2. Analysis of Effectiveness

In this sub-section we prove the non-increasing property of the updating rule presented in Eqn. (6). Considering the cost function given in Eqn. (1) and expanding $F(U_{if}^{(1)} + \Delta)$ with the second order Taylor series, we construct

$$F(U_{if}^{(1)} + \Delta) = F(U_{if}^{(1)}) + \frac{\partial F}{\partial U_{if}^{(1)}}\Delta + \frac{1}{2}\frac{\partial^2 F}{\partial^2 U_{if}^{(1)}}\Delta^2 \quad (7)$$

Recall the updating rule derived according to Eqn. (3) and (5), we have $\Delta = -\alpha_{U_{if}^{(1)}}\partial F/\partial U_{if}^{(1)}$ and

$$\begin{aligned}\alpha_{U_{if}^{(1)}} &= \frac{U_{if}^{(1)}}{\sum_{jk}(W_{ijk}\sum_{f=1}^{K}U_{if}^{(1)}U_{jf}^{(2)}U_{kf}^{(3)})U_{jf}^{(2)}U_{kf}^{(3)}} \\ &\leq \frac{U_{if}^{(1)}}{\sum_{jk}(W_{ijk}U_{if}^{(1)}U_{jf}^{(2)}U_{kf}^{(3)})U_{jf}^{(2)}U_{kf}^{(3)}} = \frac{U_{if}^{(1)}}{U_{if}^{(1)}\sum_{jk}W_{ijk}U_{jf}^{2(2)}U_{kf}^{2(3)}}\end{aligned} \quad (8)$$

From Eqn. (4), we can obtain

$$\partial^2 F/\partial^2 U_{if}^{(1)} = \sum_{jk} W_{ijk} U_{jf}^{2(2)} U_{kf}^{2(3)} \tag{9}$$

Substituting Eqn. (9) into (8), there exists $\alpha_{U_{if}^{(1)}} \leq 1/(\partial^2 F/\partial^2 U_{if}^{(1)})$. Then we have

$$\begin{aligned}
&F(U_{if}^{(1)} + \Delta) - F(U_{if}^{(1)}) \\
&\quad = -\alpha_{U_{if}^{(1)}} (\frac{\partial F}{\partial U_{if}^{(1)}})^2 + \frac{1}{2} \alpha_{U_{if}^{(1)}}^2 (\frac{\partial F}{\partial U_{if}^{(1)}})^2 \frac{\partial^2 F}{\partial^2 U_{if}^{(1)}} \\
&\quad = -\alpha_{U_{if}^{(1)}} (\frac{\partial F}{\partial U_{if}^{(1)}})^2 (1 - \frac{1}{2} \alpha_{U_{if}^{(1)}} \frac{\partial^2 F}{\partial^2 U_{if}^{(1)}}) \\
&\quad \leq -\alpha_{U_{if}^{(1)}} (\frac{\partial F}{\partial U_{if}^{(1)}})^2 (1 - \frac{1}{2}) \leq 0
\end{aligned}$$

So far, we can conclude that the iteration is non-increasing under the update rule formalized in Eqn. (6).

5. HCRF-based Activity Characterization

As shown in Fig. 2, everyday activities can be regarded as stochastic temporal processes consisting of various lengths of concept vectors. With this, the dynamic evolution of concept vector occurrences can characterize a deeper meaning of underlying, or derived, human activities if the evolution patterns can be modeled. The conditional random field (CRF) is an effective method to model temporal sequence data using an undirected graph model. While Hidden Markov Modeling (HMM) assumes all of the observations in a sequence are independent and conditional on the hidden states, CRF has no such constraints so that it allows the existence of non-local dependencies between hidden states and observations. Because the dependence is allowed in a wider range, a CRF model can flexibly adapt to dynamic sequences in which high correlations might exist between different regions. This is especially useful for modeling the kind of activities recorded by wearable lifelog cameras.

In Fig. 4 we see that due to unexpected movements of the wearer, cue concepts indicating the characteristics of activities usually appear more ran-

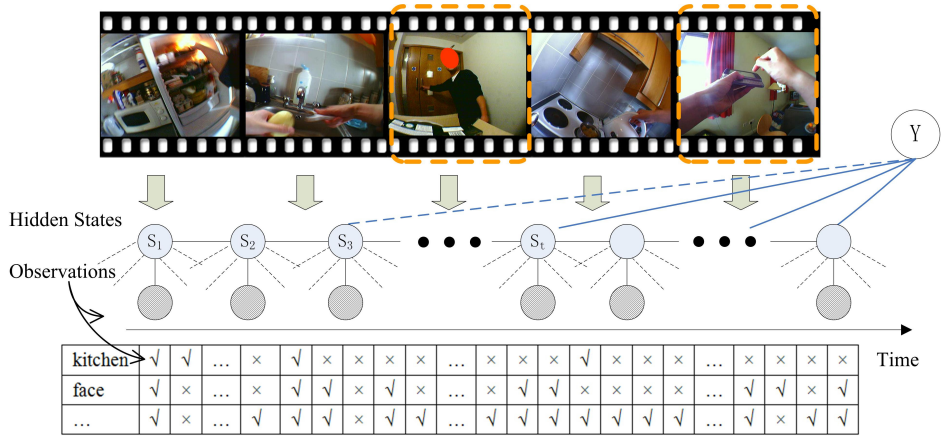


Figure 4: HCRF structure for activity modeling.

domly in the visual sequence, breaking the local consistence of concept occurrences. For example, within one recorded activity of "cooking", the lifelog subject might have talked with others or answered a phone call, as shown in Fig. 4. This introduces non-relevant concepts like "face", "mobile", etc., to which the corresponding lifelog images are highlighted by orange dashes in Fig. 4, making non-local dependencies more important in characterizing the temporal models.

We employed graphical modeling based on the HCRF [38] of activity detection to use the time-varying concept patterns from a dynamic viewpoint. By combining the outputs of WNTF-based concept detection enhancement, as discussed in Section 4, this model aims for higher-level semantic concepts which characterize underlying or derived everyday activities. Similar to HMM, HCRF also introduces a series of hidden variables $S = \{S_1, S_2, ..., S_n\}$ to each of the observations $C = \{C_1, C_2, ..., C_n\}$, i.e. the concept vector for one lifelog image, where n denotes the length of the visual lifelog steam representing one activity. Assume $\theta = \{\theta_1, \theta_2, ..., \theta_m\}$ is the model parameter, then the conditional probability can be defined as

$$P(Y, S|C, \theta) = \frac{exp\Psi(Y, S, C; \theta)}{\sum_{Y,S} exp \sum_j^m \theta_j \mathrm{F}_j(C, Y)} \tag{10}$$

where $\Psi(Y, S, C; \theta) = \sum_j^m \theta_j \mathrm{F}_j(C, Y)$ is the potential function parameterized by θ. Feature functions $\mathrm{F}_j(C, Y)$ depend on the whole sequence of concept

detection results C. By calculating the marginal probability on S [38], we can obtain

$$P(Y|C,\theta) = \sum_{S} P(Y,S|C,\theta) = \frac{\sum_S exp\Psi(Y,S,C;\theta)}{\sum_{Y,S} exp\sum_j^m \theta_j \mathrm{F}_j(C,Y)} \tag{11}$$

Therefore, the training of parameter θ can be achieved through the optimization of the objective function

$$L(\theta) = \sum_{i=1}^{n} logP(Y|C,\theta) - \frac{||\theta||^2}{2\sigma^2} \tag{12}$$

where the regularization term $\frac{||\theta||_2}{2\sigma^2}$ is applied to avoiding over-fitting and other numerical problems. The optimization can be carried out through the traditional gradient method, i.e. calculating $\theta^* = argmax_\theta L(\theta)$ through multiple iterations. In the implementation of HCRF in this paper, we employed the linear-chain structure feature functions as

$$F_j(C,Y) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, C, i) \tag{13}$$

Under this structure, each f_j depends on the whole sequence of concept detection results within one activity but is only relevant with the current and previous labels. In addition, because of the defined liner-chain structure, the objective function and its gradient can be solved in terms of the marginal distribution of hidden variables. The inference and parameter estimation can be performed by applying a belief propagation method. The described HCRF model with the above formalization can be demonstrated with the structure shown in Fig. 4.

6. Experiments and Evaluation

6.1. Experimental Setup and Datasets

For our experiments we carried out assessment of our algorithms using datasets with various levels of accuracy for concept detection. While it would be useful to have shared lifelog datasets and agreed evaluation metrics so as to allow direct comparison of techniques developed from different research teams, sharing personal lifelog data is fraught with challenges of privacy and

data ownership because lifelog data is by definition, inherenelty personal. At the present time there is simply no publicly available, shared lifelog dataset on which teams can work on shared problems, though we hope this changes soon.

To tackle the image quality problem introduced by wearers' movements despite the on-board movement sensors on the SenseCam, low-quality images are filtered out according to a fusion of the *Contrast* and *Saliency* measures which, in previous work, is shown to be effective in choosing high-quality lifelogging event representations [8, 54]. The 23 everyday activity types listed in Table 1 are applied in the evaluation for which 12,248 visual lifelog images are involved in our experiment [56]. Since there exist numerous types of activity in our everyday daily lives, those activities which have high frequency and count for more time spent would be of greatest value for applications like independent living assistance, obesity analysis, and chronic disease diagnosis. An investigation into the automatic detection of such activities would be valuable in providing insights into utilizing concept semantics in more sophisticated tasks. The criteria of time dominance, generality and high frequency are employed in selecting the 23 target activities shown in Table 1, to ensure that these activities can collectively cover most of the time spent in a typical day and are applicable to a range of individuals and age groups [55].

For this purpose, 4 people were recruited with different background demographics including a mix of older people and younger University researchers. Among these participants, one older participant is less functional in terms of capacity for household and community activities from an occupational therapist's viewpoint. The choice of these four people helps to test if our algorithm is applicable from among a group of heterogeneous subjects. All of our subjects have worn a SenseCam consecutively for more than 7 days and this allows them to get over the initial adoption and comfort issues. This guarantees a better reflection of their life patterns and visual variety of their activity samples. To address the privacy issues induced by detailed activity recordings in real life, ethical approval was first obtained from Research Ethics Committee in Dublin City University for the use of participants' SenseCam images. More details about the data sources used in this paper can be found in [56].

A detailed study of concept detector implementations is beyond the scope of this paper and motivated by [56], we simulate the detection of concepts with various accuracies based on groundtruth data. The purpose of this is

Table 1: Everyday activity types in evaluation

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

to focus on activity characterization issues rather than on concept detection. Besides this, the concept detection enhancement method proposed in this paper is independent of the specific implementation of the concept detection. Evaluation on more general detection results with different accuracy levels will further reflect its performance in improving both concept, and activity classifications. The details of the simulation are described in previous work by the authors in [56], following from work in [3].

Concept detectors for 85 everyday concepts at different accuracy levels are simulated by changing the mean of the positive class μ_1 for each concept classifier [3]. The posterior probability of concept existence is returned as the simulated concept detection output and we use this value as the original classifier confidence. For any configuration of μ_1, our evaluation is carried out with training and testing components for each run. With each parameter setting we executed 20 repeated runs and the averaged concept mean average precision (MAP) is calculated to evaluate concept detection enhancement. In Fig. 5, the averaged concept MAPs (original) over all 20 runs are plotted with different configurations of μ_1. In the generation of Fig. 5, the other three parameters are assigned with fixed values of 1.0, 1.0, 0.0 for σ_0, σ_1 and μ_0 respectively. Fig. 5 shows the improving trend for concept MAP with increasing μ_1 value and nearly perfect detection performances are achieved when $\mu_1 \geq 5.5$.

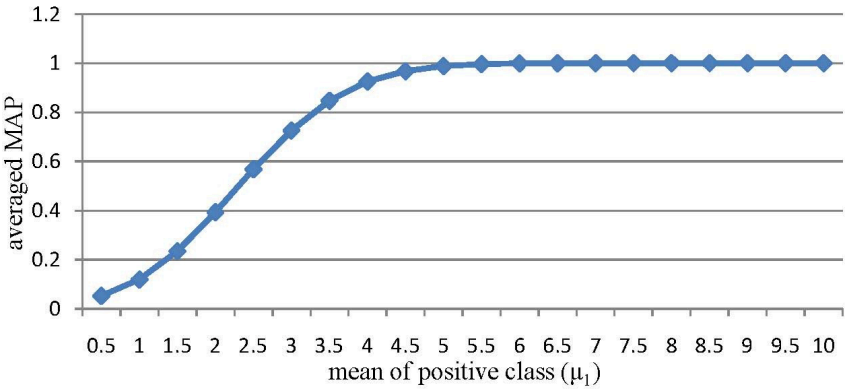


Figure 5: Averaged concept MAP with different μ_1 values.

6.2. Discussion of Results

In implementing our proposed solution, we carried out concept detection enhancement first, as described in Section 4. The HCRF-based model was then applied on the enhanced results to classify different activity categories according to the concept dynamic patterns in each visual lifelog stream. We applied a discriminative method in deciding the classification result, i.e. for each testing sequence, the likelihoods returned by HCRF models of positive and negative classes were compared and the highest was selected as the characterized activity type.

The results of our enhancement of concept detection are depicted in Fig. 6 at different μ_1 values where WNTF-based enhancement ($K = 50$, $threshold = 0.3$) significantly improves detection results when the positive class mean increases from 0.5 to 5.0. The fusion parameter in Eqn. (5) is simply set to $\alpha = 0.5$, assigning equal importance to the two tensors. The less satisfactory performance at $\mu_1 = 0.5$ is explained by the initial detection accuracy being just too low. As shown in Fig. 5, the overall MAP at $\mu_1 = 0.5$ is nearly zero. In this case, no correctly detected concept can be selected and utilised, which is actually impractical in real-world applications. When initial detection performance is good enough, as shown in Fig. 5 if $\mu_1 \geq 4.5$, there is less space to improve detection accuracy, thus the improvement is not significant at $\mu_1 \geq 4.5$.

To highlight the importance of time-awareness in concept detection enhancement, three parameter settings of $N = 1, 3, 5$ are depicted respectively in Fig. 6. When $N = 1$, the factorization problem formalized in Eqn.

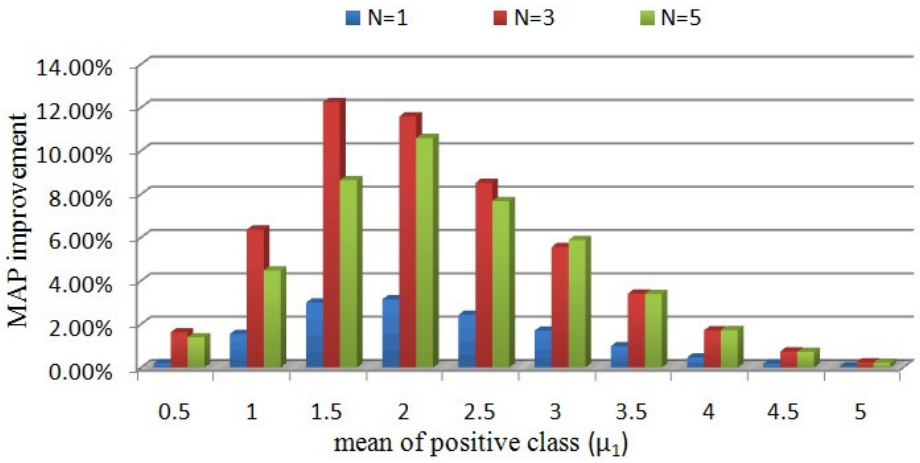


Figure 6: Enhanced concept detection by WNTF.

(1) is indeed the weighted non-negative matrix factorization (WNMF), in which the temporal information from event segmentation and the features of different events cannot be captured separately. This explains why the overall improvement at $N = 1$ is significantly out-performed by the time-aware WNTF-based method at $N = 3, 5$. This indicates that our approach can preserve local temporal constraints by introducing an extra dimension of event segment (slice as shown in Fig. 3) by representing concept detection results as a 3-way tensor.

Concept-driven activity characterization for lifelogging is a relatively new topic and the most recent report close to our work is [56] which employed a HMM to model concept dynamics and showed its effectiveness in activity modeling. In this experiment, we compare our new proposed method with HMM-based algorithm as the baseline.

To make full use of activity samples in the dataset, we decompose each set of positive samples into 50:50 ratios for training and testing respectively. While enough positive samples are necessary for the evaluation, 16 event types, each of which has more than 5 positive samples, are selected as shown in Fig. 7 for further evaluation. The activity types are shown in Table 2 with sample number and numbers of images contained. Finally, a total of 250 training samples and 250 testing samples composed of various lengths of visual lifelog streams is used for this evaluation. Because two hidden states have been tested to achieve the best overall performance in [56], to be fair,

Figure 7: Typical activity samples.

we also employed two hidden states in the experiment to train the proposed method and the baseline.

In Fig. 8, quartile comparison is shown over different original concept detection accuracies, controlled by the parameter μ_1, i.e. the mean of positive class. Each quartile is generated across 16 activity categories depicting the distribution of overall performance of 20 repeated runs. From Fig. 8 we find that the medians of the proposed method lie above the baseline at different concept detection accuracies. When the original concept detection accuracy is low, such as at $\mu_1 = 0.5$ and $\mu_1 = 1.0$, even the first quartile for the method proposed in this paper lies above the median of the baseline. At $\mu_1 = 0.5$, 75% of the scores through the proposed method are higher than the baseline. This suggests that the proposed method can better adapt to the

Table 2: Experimental data set for activity classification

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

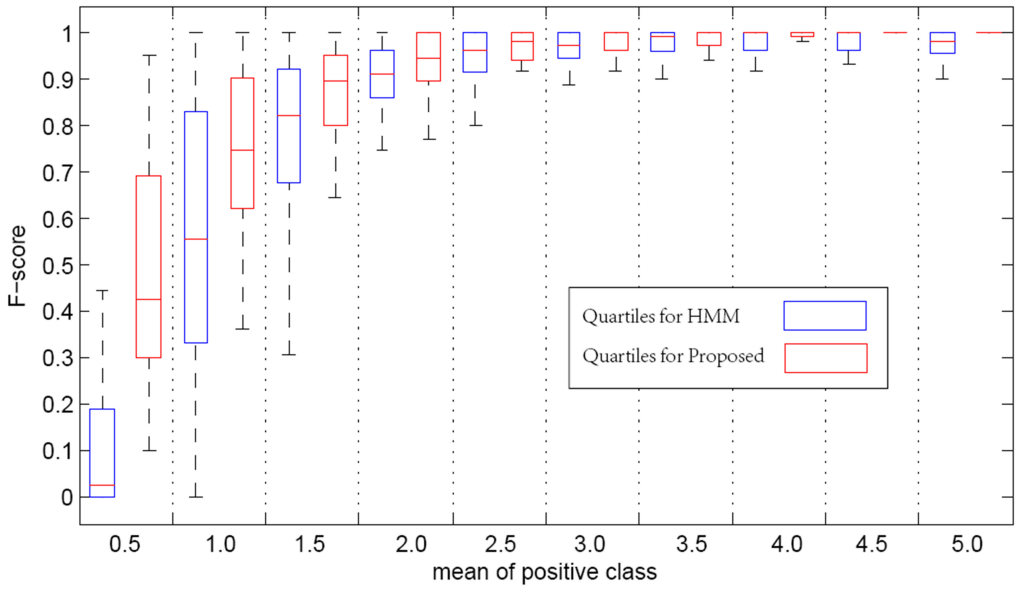


Figure 8: F-score comparison of proposed method and baseline.

low-accuracy concept detections than the baseline. Both methods perform

better when increasing the parameter μ_1 (improving concept detections) in Fig. 8. However, the proposed method shows steady performance and has lower variance than the baseline.

Table 3: Accuracy comparison on different concept detection performances (controlled by μ_1).

	$\mu_1 = 0.5$	$\mu_1 = 1.0$	$\mu_1 = 1.5$	$\mu_1 = 2.0$	$\mu_1 = 2.5$	$\mu_1 = 3.0$
Baseline	89.5%	94.7%	97.5%	98.8%	99.3%	99.5%
Proposed	94.4%	97.3%	98.5%	99.1%	99.1%	99.6%

Similar conclusions can be seen in Table 3 in which our proposed method out-performs the baseline significantly when concept detections are less satisfactory. This is consistent with Fig. 8 in that, while both methods achieve comparable accuracies at better original concept detections such as at $\mu_1 \geq 2.0$, our proposed method still shows its merits in characterizing activities at poorer concept detections.

In Table 4, pairwise activity classification F-scores are compared. From Table 4 we find that our proposed method using enhanced concept detections outperforms using non-enhanced results on most of the activities. Because more noisy concepts might be involved, the classification performances of activities like "cook", "drink", "general shopping", "talk", etc. are relatively less satisfactory for both methods. Due to the movement nature of activities like "cook", "general shopping", "talking", etc. and their wildly-varying environments, there are usually lots of concepts appearing in their image streams which are less discriminative for activity classification. For "drink", similar concept appearances like "cup", "table", "hands", etc. also introduce more misclassifications with activity "eat". In this sense, the less discriminative concepts can be termed as noisy concepts for activity classification. This shows consistency with the results as reported in [56], in which these activities have poorer F-scores even when evaluated on clean concept annotations, i.e. directly on the groundtruth. However, the method using enhanced concept detections still performs better than the non-enhanced method for most activities, showing its advantages in characterising them. For example, the "face" concept is enhanced by 9.1% and 6.7% at $\mu = 1.0$ and $\mu = 2.0$ respectively, which indicates the activity "talk" better. Similarly, the "sky" and "road/path" concepts are enhanced by 17.6% and 15.2% respectively at $\mu = 1.0$, which further enhanced the accuracy of "walk". At $\mu_1 = 2.0$, significant improvements are achieved for concepts like "monitor" (23.1%),

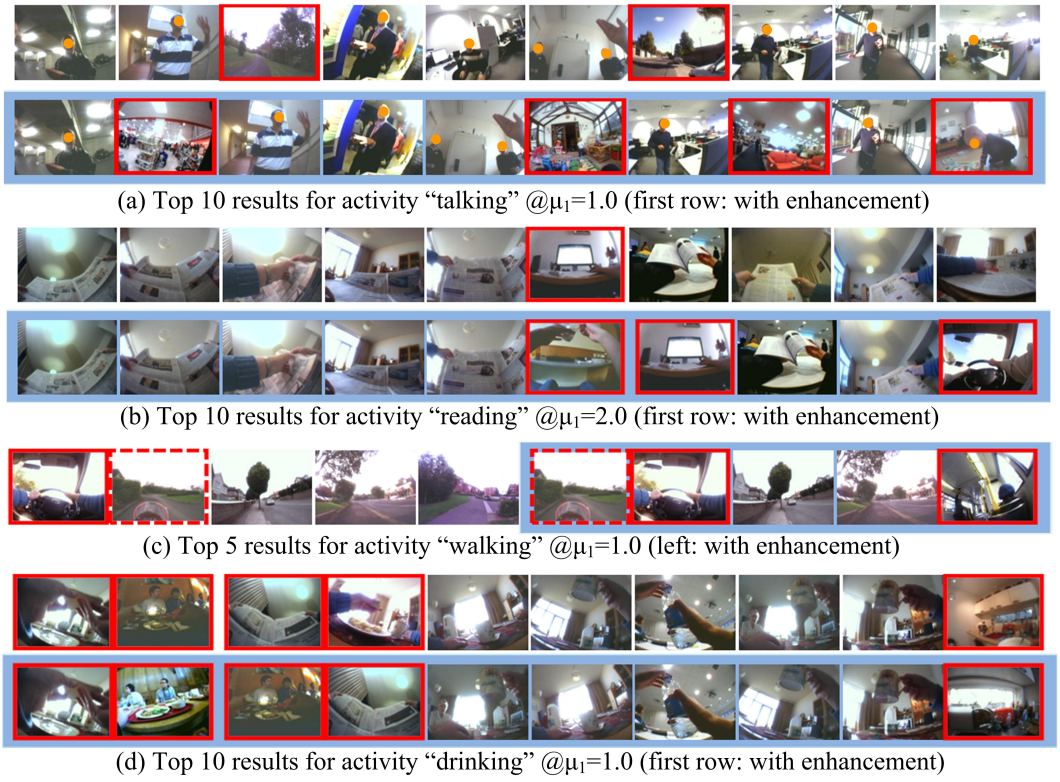
Table 4: Pairwise comparison of F-score at different concept detection performances (controlled by μ_1).

	non-enhanced		enhanced	
Activity Types	$\mu_1 = 1.0$	$\mu_1 = 2.0$	$\mu_1 = 1.0$	$\mu_1 = 2.0$
care for child	67.2%	85.0%	**67.6%**	**96.8%**
clean/tidy/wash	67.9%	88.1%	**72.6%**	**90.3%**
cook	67.6%	86.7%	65.2%	79.5%
drink	55.5%	86.0%	**60.5%**	**89.0%**
drive	98.5%	99.0%	98.4%	**99.5%**
eat	91.7%	98.9%	**93.6%**	98.6%
food shopping	74.6%	90.9%	65.2%	**92.4%**
general shopping	23.6%	58.0%	**36.1%**	**77.9%**
presentation(listen)	93.9%	95.4%	**95.1%**	**95.4%**
reading	66.8%	76.8%	**68.7%**	**94.0%**
take bus	88.9%	97.6%	**91.1%**	**98.8%**
talk	56.9%	90.6%	**68.6%**	**91.0%**
use computer	89.6%	98.2%	**92.5%**	**99.4%**
use phone	72.0%	95.6%	**73.9%**	95.5%
walk	75.0%	90.5%	**78.9%**	86.0%
watch TV	63.0%	88.8%	62.0%	**92.5%**

"screen" (15.2%), "newspaper" (17.7%), "shelf" (21.6%), etc. These can interpret an even better characterization of "watch TV", "use computer", "reading", "general shopping", etc.

Some activity recognition results are presented in Fig. 9 and visualized in returned order for each activity type. In Fig. 9 one representing frame for each sample is chosen for its visualization. The top 10 returned "talking" samples are listed in Fig. 9 (a) for our proposed method carried out on enhanced and non-enhanced concept detection results at $\mu_1 = 1.0$. Top 10 samples are both satisfactory showing the characterizing capability of the proposed method in (a). After applying the enhancement algorithm, more correct samples (first row of (a)) are returned due to better detection of concepts like "face". Similar results are also shown in Fig. 9 (b) in which more "reading" samples are correctly returned when enhancement is applied. In this case, the concept "newspaper" is enhanced and helps to recognize "reading" better. It is interesting to discover that our characterizing method can

also provide more semantics which might not be included in manual annotations. In Fig. 9 (c), one activity (highlighted in red dashes) is annotated as "child care" since "pram" appears consistently in the image stream. However, our method also returns this sample as "walking" which helps us to realize that the subject is indeed taking the baby for a walk in the garden. In activities which are semantically similar to each other such as "eating" and "drinking", concepts like "cup", "glass", "table", etc. often occur in both and introduce more difficulties in discriminating them. Taking "drinking" as an instance in 9 (d), both the enhanced and non-enhanced methods can return "eating" samples by mistake. This is because of the lack of discriminative capability of these concepts.



(a) Top 10 results for activity "talking" @μ_1=1.0 (first row: with enhancement)

(b) Top 10 results for activity "reading" @μ_1=2.0 (first row: with enhancement)

(c) Top 5 results for activity "walking" @μ_1=1.0 (left: with enhancement)

(d) Top 10 results for activity "drinking" @μ_1=1.0 (first row: with enhancement)

Figure 9: Sample results of the proposed methods. (a)–(d) represent frames for top-ranked activity samples, for "talking", "reading", "walking" and "drinking" respectively. Incorrect samples are denoted with red borders.

6.3. Computational Complexity Analysis

While the convergence property has been proven in Section 4.2, the computational efficiency of our proposed concept enhancement depends on the convergence speed of WNTF. By denoting a total number of *iter* iterations, the computational complexity is thus $O(iter \cdot NML \cdot K^3)$, where N, M and L stands for the dimensionality of input tensor T, and K denotes the rank of decomposition.

In real world applications, the lower rank K can usually achieve satisfactory results and the two dimensions of slice in Fig. 3 (M for concept number and N for image number) are much smaller than the number of slices L. Hence the computational complexity can be simplified as $O(iter \cdot L)$. In our experiments, the updating step of the approximation of $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$ *only* takes several dozens of iterations to obtain satisfactory approximation. Thus we empirically fix $iter = 100$ and it takes approximately three minutes to execute the enhancement on a conventional desktop computer.

7. Conclusions

An algorithm of WNTF used to exploit the semantics of concept re-occurrence and co-occurrence patterns is proposed and evaluated in this paper. This aims to enhance multi-concept detections for visual media captured by wearable cameras in lifelogging applications. Based on WNTF, the contextual semantics of co-occurrence and re-occurrence of semantic concepts are utilised through partial concept detection results which have better accuracies. The enhanced concept detection results are then applied to classification of everyday human activities by combining with an HCRF-based dynamic modeling algorithm. Experimental results are presented and discussed to show the efficacy of our algorithm in providing better concept and human activity detection results. Our future work is to combine the contextual semantics and temporal semantics analyzed in this paper to propose more flexible indexing approaches for wearable visual lifelogging.

Acknowledgements

This publication has resulted from research partly funded by the National Natural Science Foundation of China under Grant No. 61272231, 61472204, Beijing Key Laboratory of Networked Multimedia and by Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289.

References

[1] J.T.K. A. Jalal, N. Sharif, T.S. Kim, Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart homes, Indoor and Built Environment 22 (2013) 271–279.

[2] S.K. A. Jalal, D. Kim, A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments, Sensors 14 (2014) 11735–11759.

[3] R. Aly, D. Hiemstra, F. de Jong, P. Apers, Simulating the future of concept-based video retrieval under improved detector performance, Multimedia Tools and Applications 60 (2011) 1–29.

[4] M. Blum, A.S. Pentland, G. Tröster, InSense: Interest-based life logging, Multimedia, IEEE 13 (2006) 40 –48.

[5] M. Bukhin, M. DelGaudio, WayMarkr: Acquiring perspective through continuous documentation, in: MUM '06: Proceedings of the 5th international conference on mobile and ubiquitous multimedia, ACM, New York, NY, USA, 2006, p. 9.

[6] D. Byrne, A.R. Doherty, C.G. Snoek, G.J. Jones, A.F. Smeaton, Everyday concept detection in visual lifelogs: Validation, relationships and trends, Multimedia Tools and Applications 49 (2010) 119–144.

[7] M. Campbell, E. Haubold, S. Ebadollahi, D. Joshi, M.R. Naphade, IBM research TRECVid-2006 video retrieval system, TRECVid Online Proceedings, TRECVid 2006, 2006.

[8] A.R. Doherty, D. Byrne, A.F. Smeaton, G.J. Jones, M. Hughes, Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs, in: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, CIVR '08, ACM, New York, NY, USA, 2008, pp. 259–268.

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

[10] A.R. Doherty, A.F. Smeaton, Automatically segmenting lifelog data into events, in: WIAMIS '08: Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, IEEE Computer Society, Washington, DC, USA, 2008, pp. 20–23.

[11] A.R. Doherty, A.F. Smeaton, Automatically augmenting lifelog events using pervasively generated content from millions of people, Sensors 10 (2010) 1423–1446.

[12] J. Hamm, B. Stone, M. Belkin, S. Dennis, Automatic annotation of daily activity from smartphone-based multisensory streams, in: Mobile Computing, Applications, and Services - 4th International Conference, MobiCASE 2012, Seattle, WA, USA, October 11-12, 2012. Revised Selected Papers, pp. 328–342.

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

[14] T. Hori, K. Aizawa, Context-based video retrieval system for the lifelog applications, in: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval, MIR '03, ACM, New York, NY, USA, 2003, pp. 31–38.

[15] A.G. J. Machajdik, A. Hanbury, R. Sablatnig, Affective computing for wearable diary and lifelogging systems: An overview, in: Workshop of the Austrian Association for Pattern Recognition.

[16] A. Jalal, S. Kamal, Real-time life logging via a depth silhouette-based human activity recognition system for smart home services, in: Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on, pp. 74–80.

[17] A. Jalal, Y. Kim, D. Kim, Ridge body parts features for human pose estimation and recognition from rgb-d video data, in: Computing, Communication and Networking Technologies (ICCCNT), 2014 International Conference on, pp. 1–6.

[18] A. Jalal, M. Uddin, T.S. Kim, Depth video-based human activity recognition system using translation and scaling invariant features for life

logging at smart home, Consumer Electronics, IEEE Transactions on 58 (2012) 863–871.

[19] Y.G. Jiang, Q. Dai, J. Wang, C.W. Ngo, X. Xue, S.F. Chang, Fast semantic diffusion for large-scale context-based image and video annotation, IEEE Trans. on Image Proc. 21 (2012) 3080–3091.

[20] Y.G. Jiang, J. Wang, S.F. Chang, C.W. Ngo, Domain adaptive semantic diffusion for large scale context-based video annotation, in: Computer Vision, 2009 IEEE 12th International Conference on (ICCV), IEEE, pp. 1420–1427.

[21] Y. Jin, L. Khan, L. Wang, M. Awad, Image annotations by combining multiple evidence & WordNet, in: Proceedings of the 13th ACM Conference on Multimedia, Singapore, pp. 706–715.

[22] C. Jose, P. Goyal, P. Aggrwal, M. Varma, Local deep kernel learning for efficient non-linear svm prediction, in: Proceedings of the 30th International Conference on Machine Learning (ICML-13), pp. 486–494.

[23] P. Kelly, A.R. Doherty, A.F. Smeaton, C. Gurrin, N.E. O'Connor, The colour of life: Novel visualisations of population lifestyles, in: Proceedings of the international conference on Multimedia, MM '10, ACM, New York, NY, USA, 2010, pp. 1063–1066.

[24] L.S. Kennedy, S.F. Chang, A reranking approach for context-based concept fusion in video indexing and retrieval, in: Proceedings of the 6th ACM international conference on Image and video retrieval. (CIVR), ACM, pp. 333–340.

[25] M. Koskela, A.F. Smeaton, An empirical study of inter-concept similarities in multimedia ontologies, in: Proceedings of the 6th ACM international conference on Image and video retrieval, CIVR '07, ACM, New York, NY, USA, 2007, pp. 464–471.

[26] Q. Le, T. Sarlós, A. Smola, Fastfood – approximating kernel expansions in loglinear time, in: Proceedings of the international conference on machine learning.

[27] D. Lee, H. Seung, Learning the parts of objects by nonnegative matrix factorization, Nature 401 (1999) 788–791.

[28] H. Lee, A.F. Smeaton, N.E. O'Connor, G.J.F. Jones, M. Blighe, D. Byrne, A.R. Doherty, C. Gurrin, Constructing a SenseCam visual diary as a media process, Multimedia Systems 14 (2008) 341–349.

[29] X. Li, D. Wang, J. Li, B. Zhang, Video search in concept subspace: A text-like paradigm, in: Proceedings of the 6th ACM international conference on Image and video retrieval, CIVR '07, ACM, New York, NY, USA, 2007, pp. 603–610.

[30] S. Mann, 'WearCam' (the wearable camera): Personal imaging systems for long–term use in wearable tetherless computer–mediated reality and personal photo/videographic memory prosthesis, in: Proceedings of the 2nd IEEE International Symposium on Wearable Computers, ISWC '98, IEEE Computer Society, Washington, DC, USA, 1998, pp. 124–131.

[31] S. Mann, J. Fung, C. Aimone, A. Sehgal, D. Chen, Designing EyeTap digital eyeglasses for continuous lifelong capture and sharing of personal experiences, in: Proc. CHI 2005 Conference on Computer Human Interaction, ACM Press, Portland, Oregon, USA, 2005.

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

[33] M. Merler, B. Huang, L. Xie, G. Hua, A. Natsev, Semantic model vectors for complex video event recognition, Multimedia, IEEE Transactions on 14 (2012) 88–101.

[34] P. (MIT), `http://web.mit.edu/cron/group/house_n/placelab.html`, 2015. Online: last accessed: July. 2015.

[35] G. O'Loughlin, S. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, G.D. Warrington, Using a wearable camera to increase the accuracy of dietary analysis, Am. J. Prev. Med. 44 (2012) 297–301.

[36] G. O'Loughlin, S.J. Cullen, A. McGoldrick, S. O'Connor, R. Blain, S. O'Malley, G.D. Warrington, Using a wearable camera to increase the accuracy of dietary analysis, American Journal of Preventive Medicine 44 (2013) 297 – 301.

[37] G.J. Qi, X.S. Hua, Y. Rui, J. Tang, T. Mei, H.J. Zhang, Correlative multi-label video annotation, in: Proceedings of the 15th international conference on Multimedia, pp. 17–26.

[38] A. Quattoni, S. Wang, L.P. Morency, M. Collins, T. Darrell, Hidden conditional random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 1848–1852.

[39] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, M. Hansen, Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype, in: EmNets'07: Proceedings of the 4th workshop on Embedded networked sensors, ACM Press, Cork, Ireland, 2007, pp. 13–17.

[40] B. Safadi, G. Quénot, Evaluations of multi-learner approaches for concept indexing in video documents, in: RIAO: Adaptivity, Personalization and Fusion of Heterogeneous Information, Le Centre de Hautes Etudies Internationales d'Informatique Documentaire, pp. 88–91.

[41] A. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, K. Wood, Do life-logging technologies support memory for the past? An experimental study using SenseCam, in: Proc. CHI 2007, ACM Press, New York, NY, USA, pp. 81–90.

[42] A. Shashua, T. Hazan, Non-negative tensor factorization with applications to statistics and computer vision, in: In Proceedings of the International Conference on Machine Learning, ICML, 2005, pp. 792–799.

[43] A.R. Silva, S. Pinho, L.M. Macedo, C.J. Moulin, Benefits of SenseCam Review on Neuropsychological Test Performance , American Journal of Preventive Medicine 44 (2013) 302 – 307.

[44] A.F. Smeaton, P. Over, W. Kraaij, High level feature detection from video in TRECVid: a 5-year retrospective of achievements, in: Ajay Divakaran (Ed.), Multimedia Content Analysis, Theory and Applications, Springer, 2008, pp. 151–174.

[45] J. Smith, M. Naphade, A. Natsev, Multimedia semantic indexing using model vectors, in: Multimedia and Expo. ICME'03. International Conference on, volume 2, pp. II–445–8.

[46] C.G.M. Snoek, J.C. van Gemert, T. Gevers, B. Huurnink, D.C. Koelma, M. van Liempt, O. de Rooij, K.E.A. van de Sande, F.J. Seinstra, A.W.M. Smeulders, A.H.C. Thean, C.J. Veenman, M. Worring, The MediaMill TRECVid 2006 semantic video search engine, in: Proceedings of the TRECVid Workshop, Gaithersburg, USA.

[47] C.G.M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, M. Worring, Adding semantics to detectors for video retrieval, IEEE Transactions on Multimedia 9 (2007) 975–986.

[48] S. Song, V. Chandrasekhar, N.M. Cheung, S. Narayan, L. Li, J.H. Lim, Activity recognition in egocentric life-logging videos, in: C.V. Jawahar, S. Shan (Eds.), Computer Vision - ACCV 2014 Workshops, volume 9010 of *Lecture Notes in Computer Science*, Springer International Publishing, 2015, pp. 445–458.

[49] Y. Song, J. Tang, F. Liu, S. Yan, Body surface context: A new robust feature for action recognition from depth videos, Circuits and Systems for Video Technology, IEEE Transactions on 24 (2014) 952–964.

[50] C.C. Tan, Y.G. Jiang, C.W. Ngo, Towards textually describing complex video contents with audio-visual concept classifiers, in: Proceedings of the 19th ACM International Conference on Multimedia, MM '11, ACM, New York, NY, USA, 2011, pp. 655–658.

[51] C. Wang, F. Jing, L. Zhang, H.J. Zhang, Image annotation refinement using random walk with restarts, in: Proc. Annual ACM international conference on Multimedia, pp. 647–650.

[52] C. Wang, F. Jing, L. Zhang, H.J. Zhang, Content-based image annotation refinement, in: Computer Vision and Pattern Recognition, 2007. (CVPR). IEEE Conference on, pp. 1–8.

[53] P. Wang, A.F. Smeaton, Aggregating semantic concepts for event representation in lifelogging, in: Proceedings of the International Workshop on Semantic Web Information Management, SWIM '11, ACM, New York, NY, USA, 2011, pp. 8:1–8:6.

[54] P. Wang, A.F. Smeaton, Aggregating semantic concepts for event representation in lifelogging, in: Proceedings of the International Work-

shop on Semantic Web Information Management, SWIM '11, ACM, New York, NY, USA, 2011, pp. 8:1–8:6.

[55] P. Wang, A.F. Smeaton, Semantics-based selection of everyday concepts in visual lifelogging, International Journal of Multimedia Information Retrieval 1 (2012) 87–101.

[56] P. Wang, A.F. Smeaton, Using visual lifelogs to automatically characterize everyday activities, Information Sciences 230 (2013) 147–161.

[57] P. Wang, A.F. Smeaton, C. Gurrin, Factorizing time-aware multi-way tensors for enhancing semantic wearable sensing, in: X. He, S. Luo, D. Tao, C. Xu, J. Yang, M. Hasan (Eds.), MultiMedia Modeling, volume 8935 of *Lecture Notes in Computer Science*, Springer International Publishing, 2015, pp. 571–582.

[58] Y. Wu, B. Tseng, J. Smith, Ontology-based multi-classification learning for video concept detection, in: Multimedia and Expo, 2004. (ICME). IEEE Intnl. Conf., pp. 1003–1006, Vol.2.