Factorizing Time-Aware Multi-Way Tensors for Enhancing Semantic Wearable Sensing

Peng Wang*, Alan F. Smeaton, and Cathal Gurrin

National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University
pengwangnudt@sina.com
Insight Centre for Data Analytics
Dublin City University, Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie, cathal.gurrin@computing.dcu.ie

Abstract. Automatic concept detection is a crucial aspect of automatically indexing unstructured multimedia archives. However, the current prevalence of one-per-class detectors neglect inherent concept relationships and operate in isolation. This is insufficient when analyzing content gathered from wearable visual sensing, in which concepts occur with high diversity and with correlation depending on context. This paper presents a method to enhance concept detection results by constructing and factorizing a multi-way concept detection tensor in a time-aware manner. We derived a weighted non-negative tensor factorization algorithm and applied it to model concepts' temporal occurrence patterns and show how it boosts overall detection performance. The potential of our method is demonstrated on lifelog datasets with varying levels of original concept detection accuracies.

Keywords: visual lifelogging, concept detection, NTF, concept semantics, wearable sensing

1 Introduction

With the maturity of lightweight sensors and computing devices, and more recently the emergence of unobtrusive wearable visual sensing devices like Google Glass or Microsoft's SenseCam, the creation of large volumes of personal, first-person visual media archives for quantified-self applications has become feasible. Visual lifelogging is the term used to describe one class of personal sensing and digital recording of all our everyday behaviour which employs wearable cameras to capture image or video of everyday activities [1].

To manage what is in effect a new form of multimedia, the lifelog, state-of-the-art techniques suggest that we use statistical mapping from low-level visual

* This research work was part-funded by the National Natural Science Foundation of China under Grant No. 61272231, 61472204, Beijing Key Laboratory of Networked Multimedia and Science Foundation Ireland under grant SFI/12/RC/2289

features to semantic concepts which are more appropriate to users' understanding of their lifelogs. According to the TRECVid benchmark, acceptable results in mapping low level features to semantic concepts have been achieved already, particularly for concepts for which there exists enough annotated training data [2]. However, unlike most other kinds of multimedia content, a wide range of semantic concepts will usually appear in visual lifelogs because of the wide variety of activities that people usually engage in and which are subsequently logged and recorded. In addition, due to the wearers' movements while capturing a visual lifelog, images captured within the same event or activity may have significant perceptual differences as, for example, users will turn around and face a window while still being in the same room. This poses many challenges for the organisation of wearable visual lifelogs which is essential if lifelogs are to be used to good effect.

In addition to visual media, a rich pool of information can be collected in wearable sensing by individuals to record their own activities and this can be used to build applications that enhance their quality of life in many ways including productivity, health monitoring and wellness, safety and security, social interactions, leisure and more. However, the raw lifelog data has comparatively little metadata and so performing content-based operations on the lifelog is problematic, especially as the archives become larger. Accurately structuring a lifelog into events [3] is considered crucial in managing visual logs for various applications, and the identification of events and event boundaries [4] is normally the first step in processing lifelogs. However, this alone doesn't offer a complete solution because we need to know what the contents of events actually are and how they relate to each other. Therefore, the focus in research has shifted towards mining deeper meanings from visual lifelogs and lifelog events i.e. determining the semantics reflected in lifelogs.

Concepts express the semantics of media in a useful way and are usually automatically detected by providing a meaningful link between low-level features like colours and textures, and high-level semantics. In [5], the semantic indexing method has shown potential for relating low-level visual features to high-level semantic concepts (such as indoors, outdoors, people, buildings, etc.) for visual lifelogs using supervised machine learning techniques. This is then applied in [6] to learn lifestyle traits from lifelogs collected by different users, based on the automatically detected everyday concepts. The accuracy of a concept detector/classifier is an important factor in the provision of satisfactory solutions to indexing visual media and it is also widely accepted that detection accuracy can be improved if concept correlation can be utilised. The utilization of correlation in multi-concept detection falls into two main categories: multi-label training and detection refinement/adjusting. A typical multi-label training method is presented in [11], in which concept correlations are modeled in the classification model using Gibbs random fields. Since all concepts are learned from one integrated model, the direct shortcoming is the lack of flexibility, which means the learning stage needs to be repeated when concept lexicon is changed. Because detection scores obtained by specific binary detectors allow independent and possibly specialized classification techniques to be leveraged for each concept [14], detection refinement using post processing attracts much research interest based on utilising concept correlations inferred from preconstructed knowledge [12, 15] or annotation sets [16–18]. These methods highly depend on external knowledge such as WordNet or the training data. When concepts do not exist in the lexicon ontology or extra annotation sets are insufficient for correlation learning (limited size of corpus or sparse annotations), these methods can not adapt to these situations and obtain equally good results. In [19], a semantic enhancement method is proposed for lifelogging based on weighted none-negative matrix factorization (WNMF), but the temporal semantics can not been applied in this model.

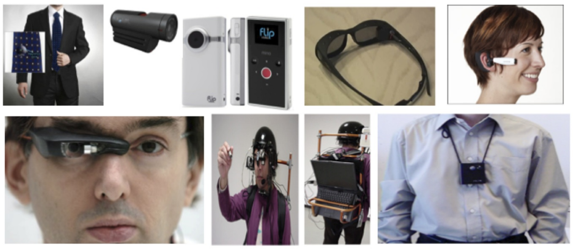

Fig. 1: A variety of wearable visual lifelog devices through the ages including SenseCam (bottom right).

In this paper, we propose an enhancement to concept detection by using inherent inter-concept correlations. Based on the assumption that the scores from the initial detectors are reasonably usable for some concepts similar as in [16], our method exempts from using any extra annotation sets and includes concept detection results as the only input. To evaluate the effectiveness of our approach to enhancing concept detection, we employed SenseCam (shown in Figure 1) as a wearable device to log details of users' lives. SenseCam has a lightweight passive camera with several built-in sensors which captures the view of the wearer with its fisheye lens. By default, images are taken at the rate of about one every 50 seconds while the on-board sensors can help to trigger the capture of pictures when sudden changes are detected in the environment of the wearer.

2 Overview of Problem and Solution

We define the research problem as follows: given particular streams of everyday activities divided into discrete events with consecutive images each of which has some concepts detected, the task is to use each concepts' contextual semantics, embedded in the detection results, to improve the overall detection performance. We assume a lexicon of concepts L. Let $\{E_1, E_2, ..., E_n\}$ be the set of event streams in the dataset. Event E_i is represented by successive images $I^{(i)} =$

$\{Im_1^{(i)}, Im_2^{(i)}, ..., Im_k^{(i)}\}$. Each image $Im_j^{(i)}$ might have several concepts detected. We assume the concepts appearing in image $Im_j^{(i)}$ are represented as a confidence vector $C_j^{(i)} = \{c_{j1}^{(i)}, c_{j2}^{(i)} ... c_{jM}^{(i)}\}$ for M concepts. The whole set of SenseCam images can be denoted as $I = \{I^{(1)}, I^{(2)}, ..., I^{(n)}\}$ which has dimension $\sum_{i=1}^{n} k_i$, where k_i is the number of images in each event E_i.

Concatenating confidence vectors from all SenseCam images represents detection results as a 2-dimensional matrix, however this loses information from event segmentation and the features of different events are not captured separately. To utilise the temporal features reflected in different events, a tensor is employed to formalize the above problem given its merit in representing the structure of multidimensional data more naturally than matrices. The algorithm for enhancing concept detection proposed in this paper requires a nonnegative tensor factorization (NTF) approach to capture latent feature structure. By introducing a new dimension, NTF can preserve and model temporal characteristics of each event and avoid significant information loss.

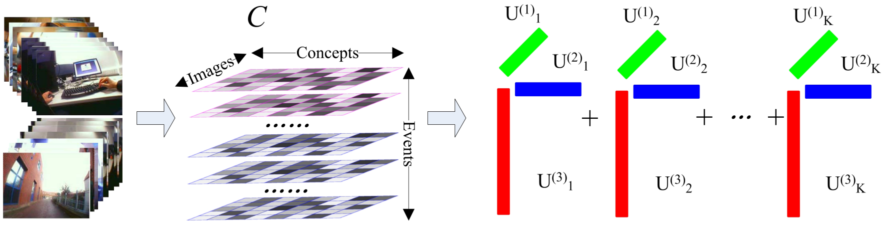

Fig. 2: NTF-based concept detection enhancement framework.

The procedure for concept tensor construction and factorization is illustrated in Figure 2. As shown, our approach treats the concept detection results in a natural 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. In Figure 2, we use different colors of slices to show that they are the segments from different events. Meanwhile, the confidences of concept existences in each slice are represented by various gray levels. 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.

Assume each slice is a segment of N SenseCam images, each of which is represented by a vector of M concept detection confidences (i.e. concept vectors). The constructed concept detection tensor C has the dimensionality of $N \times M \times L$ for events with L slices in total. The task now is to modify the $N \times M \times L$ dimensional tensor C in order to keep consistency with the underlying contextual pattern of concepts. The factorization of weighted non-negative tensor C and the concept detection enhancement based on this WNTF method is now described.

3 Time-Aware Concept Detection Enhancement

3.1 Weighted Non-Negative Tensor Factorization (WNTF)

As we can see from Section 2, the confidence tensor C has a dimensionality of $N \times M \times L$ which consists of N neighborhood SenseCam images, M semantic concepts and L time intervals. The task of WNTF is to find the latent features to represent the three components of confidence tensor C. The tensor can then be approximated by the Tucker Decomposition (TD) [8] as

$$C \approx G \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}$$

where $G \in \mathcal{R}^{R\times S\times T}$, $U^{(1)} \in \mathcal{R}^{N\times R}$, $U^{(2)} \in \mathcal{R}^{M\times S}$ and $U^{(3)} \in \mathcal{R}^{L\times T}$. The operator $\times_i (i = 1, 2, 3)$ denotes the tensor-matrix multiplication operators with the subscript $_i$ specifying which dimension of the tensor is multiplied with the given matrix. In Tucker Decomposition, the high-order tensor is factorized into a core tensor G and a factor matrix $U^{(i)}$ along each mode i [9]. In the TD model, each element in C is approximated by

$$\hat{C}_{ijk} = \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T} G_{rst} U^{(1)}_{ir} U^{(2)}_{js} U^{(3)}_{kt}$$

As a particular case of the general Tucker Decomposition, the Canonical Decomposition (CD) [10] is derived from the TD model by constraining that each factor matrix has the same number of columns, i.e., the length of latent features has a fixed value of K. By setting G as a diagonal tensor

$$G_{ijk} = \begin{cases} 1, \text{ if } i = j = k \\ 0, \text{ else} \end{cases}$$

the CD model simplifies the approximation of tensor C as a sum of 3-fold outer-products with rank-K decomposition $\hat{C} = \sum_{f=1}^{K} U^{(1)}_{\cdot f} \otimes U^{(2)}_{\cdot f} \otimes U^{(3)}_{\cdot f}$, which means that each element $\hat{C}_{ijk} = \sum_{f=1}^{K} U^{(1)}_{if} U^{(2)}_{jf} U^{(3)}_{kf}$.

The CD approximation factorization defined above can be solved by optimizing the cost function defined to qualify the quality of the approximation. Different forms of cost function and corresponding optimization can be applied to this problem. Euclidian distance can be used to define the cost function, which has the form of $F = \frac{1}{2}\|C - \hat{C}\|^2_F$. However, in factorizing the confidence tensor, the weighted measure is more suitable since detection performance differs due to the characteristics of concepts and quality of the training set. To distinguish the contribution of different concept detectors to the cost function, the weighted cost function is employed as

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

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. The nonnegative constraints guarantees each component described by $U^{(1)}$, $U^{(2)}$, $U^{(3)}$ are additively combined.

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

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

To solve the partial differential $\partial F/\partial U^{(t)}$, we can rewrite Equation (1) as

$$F = \frac{1}{2} < C - \hat{C}, C - \hat{C} >_W = \frac{1}{2} < C - \sum_{f=1}^{K} \otimes_{t=1}^{3} U_{\cdot f}^{(t)}, C - \sum_{f=1}^{K} \otimes_{t=1}^{3} U_{\cdot f}^{(t)} >_W$$

where $< X, Y >$ denotes the inner product of two 3-way tensors [20] which is defined as $< X, Y >= \sum_{ijk} x_{ijk} y_{ijk}$. Hence we conduct the derivative

$$dF = \frac{1}{2} d < C - \hat{C}, C - \hat{C} >_W = < W \circ (C - \hat{C}), -d(\sum_{f=1}^{K} \otimes_{t=1}^{3} U_{\cdot f}^{(t)}) > \tag{3}$$

Without losing generality, we focus on the update of the fth column in $U^{(1)}$ in the following derivation procedure and the update rule for other columns and matrices can be obtained in a similar manner. By taking the differential with respect to $U_{\cdot f}^{(1)}$, we can obtain the derivative of Equation (3) as

$$\begin{aligned} dF(U_f^{(1)}) &= < W \circ (C - \hat{C}), -d(U_{\cdot f}^{(1)}) \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)} > \\ &= < W \circ \hat{C}, d(U_{\cdot f}^{(1)}) \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)} > \\ &- < W \circ C, d(U_{\cdot f}^{(1)}) \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)} > \end{aligned}$$

Hence the differential with respect to an element $U_{if}^{(1)}$ can be represented as

$$\begin{aligned} \partial F/\partial U_{if}^{(1)} &= < W \circ \hat{C}, e_i \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)} > - < W \circ C, e_i \otimes U_{\cdot f}^{(2)} \otimes U_{\cdot f}^{(3)} > \\ &= \sum_{jk} (W \circ \hat{C})_{ijk} U_{jf}^{(2)} U_{kf}^{(3)} - \sum_{jk} (W \circ C)_{ijk} U_{jf}^{(2)} U_{kf}^{(3)} \end{aligned}$$

where e_i is the ith column of the identity matrix and has the same dimension as $U_{\cdot f}^{(1)}$. By employing $\alpha_{U^{(1)}}$ as the form $\alpha_{U_{if}^{(1)}} = U_{if}^{(1)} / \sum_{jk} (W \circ \hat{C})_{ijk} U_{jf}^{(2)} U_{kf}^{(3)}$, where / denotes element-wise division, and substituting into Equation (2), we obtain the multiplicative updating rule [21] as

$$U_{if}^{(1)} \leftarrow U_{if}^{(1)} \frac{\sum_{jk} (W \circ C)_{ijk} U_{jf}^{(2)} U_{kf}^{(3)}}{\sum_{jk} (W \circ \hat{C})_{ijk} U_{jf}^{(2)} U_{kf}^{(3)}}$$

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

3.2 WNTF-Based Concept Detection Enhancement

To obtain a reconstruction of the underlying semantic structure that we can mine for co-occurrences and so enhance raw concept detection performance, the weights must be set in terms of concept accuracy. Because each confidence value C_{ijk} in tensor C denotes the probability of concept C_j occurring in the image, estimating the existence of C_j is more likely to be correct when C_{ijk} is high enough. Under this premise [16], we used the concept detection enhancement as in Algorithm 1:

Algorithm 1: WNTF-based detection enhancement

```
Input:
$C = (C_{ijk})_{N \times M \times L}$: original confidence tensor, $threshold$
Output:
$C_{new} \in \Re_{N \times M \times L}$: adjusted confidence tensor for $C$
Data:
$W \in \Re_{N \times M \times L}$: weight tensor
$U^{(1)} \in \Re_{N \times K}, U^{(2)} \in \Re_{M \times K}, U^{(3)} \in \Re_{L \times K}$
begin
    Normalize $C$ at each concept slice:
    $C(:, j, :) = normalize(C(:, j, :)), 1 \leq j \leq M$;
    Initialized $U^{(1)}, U^{(2)}, U^{(3)}$ randomly with small numbers;
    for each $C_{ijk}$ in $C$ do
        $C'_{ijk} = C_{ijk}, W_{ijk} = 1$ if $C_{ijk} \geq threshold$;
        $C'_{ijk} = 0, W_{ijk} = w, w \in (0, 1)$; Otherwise;
    repeat
        $U^{(1)}_{if} \leftarrow U^{(1)}_{if} \sum_{jk} (W \circ C')_{ijk} U^{(2)}_{jf} U^{(3)}_{kf} / \sum_{jk} (W \circ \hat{C}')_{ijk} U^{(2)}_{jf} U^{(3)}_{kf}$
        $U^{(2)}_{jf} \leftarrow U^{(2)}_{jf} \sum_{ik} (W \circ C')_{ijk} U^{(1)}_{if} U^{(3)}_{kf} / \sum_{ik} (W \circ \hat{C}')_{ijk} U^{(1)}_{if} U^{(3)}_{kf}$
        $U^{(3)}_{kf} \leftarrow U^{(3)}_{kf} \sum_{ij} (W \circ C')_{ijk} U^{(1)}_{if} U^{(2)}_{jf} / \sum_{ij} (W \circ \hat{C}')_{ijk} U^{(1)}_{if} U^{(2)}_{jf}$
    until $Converges$;
    for each $C'_{ijk} \in [C']_0$ do
        $C'_{ijk} = \sum_{f=1}^{K} U^{(1)}_{if} U^{(2)}_{jf} U^{(3)}_{kf}$
Return $C_{new} = [average(C'_{ijk}, C_{ijk})]_{N \times M \times L}$;
```

Firstly, each concept-oriented slice $C(:, j, :)$ of tensor C is normalized at $Max-Min$ scale [14] for each specific concept j, which is indeed a lateral slice in the tensor visualized by Figure 2. This is then followed by constructing a new sparse tensor C' by thresholding C, whose element is

$$C'_{ijk} = \begin{cases} C_{ijk}, & \text{if } C_{ijk} \geq threshold; \\ 0, & \text{otherwise.} \end{cases}$$

The rationale for this is to retain elements with high confidence as "seeds" and use the contextual information modeled by non-negative tensor factorization to predict other concepts in correlation with these seed concepts. A sparse confidence tensor C' is achieved and we denote the non-zero element set in C' as

$[C']_+$. Meanwhile, the set $[C']_0$ can be used to denote zero elements in C' which need to be estimated from $[C']_+$. C' is then factorized using the updating algorithm described in Section 3.1. This involves the iterative optimization of the cost function defined in Equation (1). In the optimization step, we configure the settings of weights as $W_{ijk} = 1$ if $C'_{ijk} \in [C']_+$, otherwise $W_{ijk} \in (0,1)$. In this step, the component matrices of $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$ are returned as an estimate of the contextual structure of C'.

Finally, the approximation of elements in $[C']_0$ can be calculated using the refactorized features as $\sum_{f=1}^{K} U^{(1)}_{if} U^{(2)}_{jf} U^{(3)}_{kf}$, in which each component is the latent factor learned from the sparse tensor C'. The new confidence values for elements in $[C']_0$ form an estimate of concept detection to adjust the original detection result by averaging the original confidence and the new estimated value.

4 Results and Discussion

4.1 Experimental Setup and Dataset

To assess the performance of our algorithm, we used a set of 85 everyday concepts and a dataset including event samples of 23 activity types collected from 4 SenseCam wearers consisting of 12,248 SenseCam images [13]. Concept detectors with different accuracy levels were simulated and the metrics of AP and MAP were calculated for concepts based on a manual groundtruth. Different concept detection accuracies were provided in the dataset by varying the mean of the positive class μ_1 in the range [0.5...10]. The details of simulation are described in [13], following on from the work by Aly in [7]. For each setting of parameters, we executed 20 repeated runs to avoid random performance and the averaged concept AP and MAP were both calculated. The accuracy of the detection of original concepts is simulated with various accuracy levels and the MAPs are shown in Table 1 (first row) with the increased values of simulation parameter μ_1. The rationale for this is to test the performance of our algorithm at different concept detection accuracies. WNTF-based enhancement is carried out as described in Section 3 with concept detection confidence as the only input.

4.2 Detection Enhancement Analysis

Since averaging MAP over different detection accuracies is meaningless, pairwise comparison is depicted in Table 1 at different μ_1 values where detection enhancement ($K = 50$, $threshold = 0.3$) is applied. As shown by the improvement in Table 1, our algorithm can self-learn the contextual semantics of concepts and enhance the overall detection performance for various original detection accuracy levels. The highest overall improvement of 10.59% is achieved at $\mu_1 = 2.0$ when the original detection performance is neither too low nor too high. The improvement is shown to be significant and robust at various original detection accuracy levels in Table 1.

Table 1: Improved concept detections for various original accuracies.

Value of μ_1	**0.5**	**1.0**	**1.5**	**2.0**	**2.5**	**3.0**	**4.0**	**5.0**
Original MAP	0.0946	0.1570	0.2645	0.4124	0.5797	0.7313	0.9251	0.9891
Adjusted MAP	0.0959	0.1640	0.2874	0.4560	0.6242	0.7744	0.9410	0.9912
Improvement	1.40%	4.48%	8.65%	**10.59%**	7.69%	5.89%	1.72%	0.21%

The less significant performance of our enhancement approach at $\mu_1 = 5.0$ makes sense as the initial detection accuracy is good enough. In this case, there is no space to improve detection accuracy, which is also the case when $\mu_1 = 4.0$ at which the original MAP has already reached 0.9. However, our approach can still enhance detection results with an improvement of 1.72% at $\mu_1 = 4.0$. On the other hand, when the original detection accuracy is too low, as shown in Table 1 at $\mu_1 = 0.5$, low accuracy detected elements can be selected and treated as "seed" candidates in our algorithm. Though this is an extreme, which is impractical in real world applications, our approach still works well with the average improvement of 1.40% achieved.

In many lifelog application scenarios, concept detection confidences need to be binarized to decide the existence or absence of concepts, instead of using the raw concept detection confidence values. Figures 3 and 4 illustrate the F-score, Recall-Precision improvement at different binarization levels $threshold_{bin}$, after applying our enhancement algorithm, taking the two concepts 'inside bus' and 'building' as instances. To consider the role of different filtering values of $threshold$ in Section 3.2, we assign $threshold = 0.5$ and $threshold = 0.8$ in Figure 3 and Figure 4 respectively. As shown by these two figures, the curves for two concepts are both enhanced. Since we use $threshold = 0.5$ in implementing the WNTF-based method in Figure 3, a large proportion of the adjusted concept detection confidences are below this threshold value. Hence the enhancements are significant for the parts of curves when the binarization $threshold_{bin} < 0.5$. In this case, if we choose the binarizing threshold at higher values such as $threshold_{bin} \geq 0.5$, the use of WNTF will affect the result less significantly because most of the adjusted confidences are less than $threshold_{bin}$ and the corresponding concepts are still decided not to be present in the SenseCam images. Meanwhile, if we choose a higher value of $threshold = 0.8$, a larger range of enhancement for Recall and Precision can be achieved as shown by the curves in Figure 4.

Our algorithm has the advantage of enhancing a large number of concepts as demonstrated in Figure 5. In Figure 5, the performances of WNTF-based ($K = 50$) and WNMF-based [19] methods are compared across all 85 concept APs using the same $threshold$. The detection of around 60 concepts are improved by our algorithm at $\mu_1 = 1.5$. In [19], the advantage of WNMF-based method has been demonstrated against ontological method for lifelogging concept enhancement. However, by utilising the temporal features, the WNTF-based method is more effective and the overall improvement is significant across all 85 concepts.

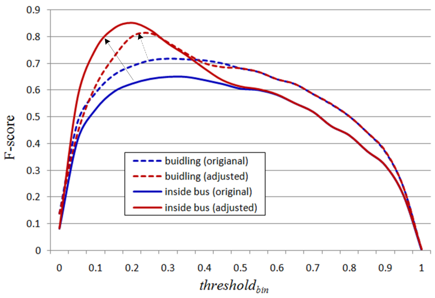

Fig. 3: F-score enhanced at $threshold =$ 0.5 (for 'inside bus' and 'building').

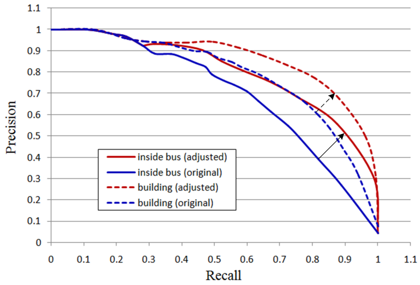

Fig. 4: Recall-precision curve enhanced at $threshold = 0.8$ (for 'inside bus' and 'building').

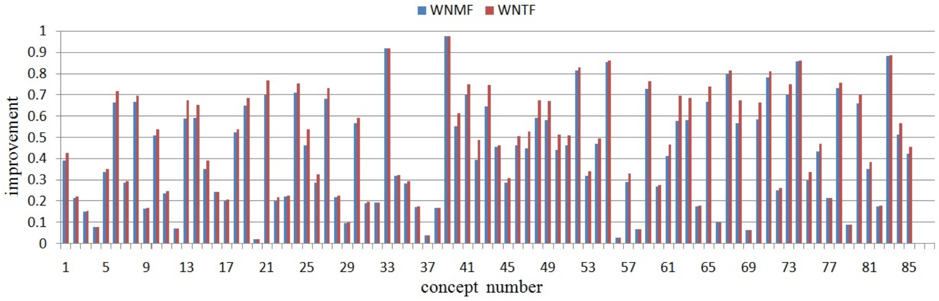

Fig. 5: Improvement comparison over all concepts.

4.3 Impact of Parameters

The impact of parameters on enhancement performance is demonstrated in Figures 6 and 7, in which improvement is depicted at two different concept detection accuracies, determined by $\mu_1 = 1.5$ and $\mu_1 = 2.5$ respectively. In Figure 6, results for all settings of $K \in [10, ..., 80]$ and $threshold \in [0.1, ..., 0.9]$ are shown. All cases in Figure 6 are achieved by executing the algorithm in 20 runs and the averaged MAP improvement across all 85 concepts are obtained. We notice the robustness of the WNTF-based enhancement algorithm through the improvements achieved over different configurations of K and $threshold$.

As shown in Figure 6, detection performance is improved in most cases when the value of $threshold$ is not very high. The reason is because when $threshold$ is chosen as too high, there will be fewer correct concept detection results chosen, hence the potential for overall performance improvement is lessened. As shown in Figure 6, the best overall performances are achieved when $threshold = 0.3$ for $\mu_1 = 1.5$. In Figure 7, for which the original MAP is better as shown in Table 1, more correctly detected concepts can be used when higher $threshold$ is chosen to give better estimates on the others. That is why $threshold = 0.5$ achieves the

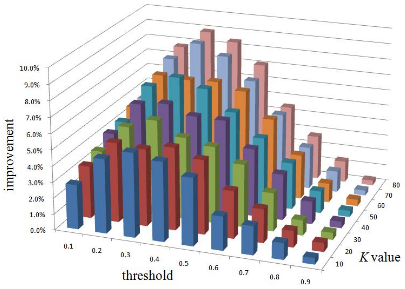

Fig. 6: MAP improvement with various parameter configures ($\mu_1 = 1.5$).

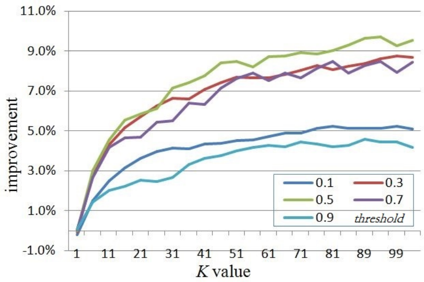

Fig. 7: Impact of feature number K ($\mu_1 = 2.5$).

best performance for Figure 7. The choice of “noisy” concepts can also degrade the improvement, as depicted when $threshold$ is small, say, $threshold = 0.1$ in both figures. In these cases, erroneous detection results are likely to be chosen to $[C']_+$ in the thresholding procedure which contaminates performance.

The impact of selected latent features is shown in Figure 7 in which the improvement in detection for different $threshold$ values are depicted across different K values. With the increase in K, performance improves gradually and converges at stable values. For poorly-chosen $threshold$ values such as 0.1 and 0.9, the performance converges earlier, which reduces the potential for improvement if we increase the number of features. This implies that higher dimensionality is necessary to characterize the semantic features of concepts when more correct concept detection results are selected as “seeds” in the enhancement. For all settings of $threshold$, the performance keeps increasing and usually achieves satisfactory enhancement when about 50 latent features are selected.

5 Conclusions and Future Work

We present an algorithm to improve performance of semantic concept detection for wearable visual sensing. Based on non-negative tensor factorization, the algorithm models concept appearance patterns through partial concept detection results, which have better accuracy. For this purpose, we derived a weighted factorization method for updating latent features representing the structure of a multi-way confidence tensor. Based on this weighted nonnegative tensor factorization, local temporal constraints in each event segment are retained and reflected for the time-aware enhancement which uses the concept co-occurrence and re-occurrence patterns. The confidences of less accurate concept detections are then estimated and adjusted to enhance performance of overall concept detection. This method has been evaluated in experiments on datasets with various original detection accuracies. Since the factorization of time-aware WNTF also models the temporal structure of events, the application of this approach to event structuring and detection is a promising suggestion for future work.

References

1. Gurrin, C., Smeaton, A. F., Doherty, A.: LifeLogging: personal big data. Foundations and Trends in Information Retrieval **8**(1), pp. 1–127, (2014)
2. A. Smeaton, P. Over, and 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.
3. Doherty, A.R., Pauly-Takacs, K., Caprani, N., Gurrin, C., Moulin, C.J.A., O'Connor, N.E., Smeaton, A.F.: Experiences of aiding autobiographical memory using the SenseCam. Human-Computer Interaction **27**(1-2), 151–174 (2012)
4. Doherty, A.R., Smeaton, A.F.: Automatically segmenting lifelog data into events. In: WIAMIS'08, pp. 20–23. IEEE Computer Society, Washington, DC, USA (2008)
5. Byrne, D., Doherty, A.R., Snoek, C.G.M., Jones, G.J.F., Smeaton, A.F.: Everyday concept detection in visual lifelogs: validation, relationships and trends. Multimedia Tools Appl. **49**(1), 119–144 (2010)
6. Doherty, A.R., Caprani, N., O'Conaire, C., Kalnikaite, V., Gurrin, C., O'Connor, N.E., Smeaton, A.F.: Passively recognising human activities through lifelogging. Computers in Human Behavior **27**(5), 1948–1958 (2011)
7. Aly, R., Hiemstra, D., de Jong, F., Apers, P.: Simulating the future of concept-based video retrieval under improved detector performance. Multimedia Tools and Applications **60**(1), 1–29 (2011)
8. Tamara G. Kolda, Brett W. Bader: Tensor decompositions and applications. SIAM Review **51**(3), 455–500 (2009)
9. Tamara G. Kolda: Multilinear operators for higher-order decompositions. Tech. Report SAND2006-2081, (2006)
10. Rendle, Steffen, Schmidt-Thieme, Lars: pairwise interaction tensor factorization for personalized tag recommendation. In: WSDM'10, pp. 81–90. (2010)
11. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J.: Correlative multi-label video annotation. In: ACM MM'07, pp. 17–26. (2007)
12. Wu, Y., Tseng, B., Smith, J.: Ontology-based multi-classification learning for video concept detection. In: ICME'04, pp. 1003–1006. vol. 2. (2004)
13. Wang, P., Smeaton, A.F.: Using visual lifelogs to automatically characterise everyday activities. Information Sciences **230**, 147–161 (2013)
14. Smith, J. R., Naphade, M., Natsev, A.: Multimedia semantic indexing using model vectors. In: ICME'03, pp. 445–448. vol. 2. (2003)
15. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & WordNet. In: ACM MM'05, pp. 706–715. (2005)
16. Kennedy, L. S., Chang, S. F.: A reranking approach for context-based concept fusion in video indexing and retrieval. In: CIVR '07, pp. 333–340. (2007)
17. Wang, C. H., Jing, F., Zhang, L., Zhang, H. J.: Image annotation refinement using random walk with restarts. In: ACM MM'06, pp. 647–650. (2006)
18. Wang, C. H., Jing, F., Zhang, L., Zhang, H. J.: Content-based image annotation refinement. In: CVPR'07, pp. 1-8. (2007)
19. Wang, P., Smeaton, A.F., Zhang, Y. C., et al.: Enhancing the detection of concepts for visual lifelogs using contexts instead of ontologies. In: ICMEW, pp. 1-6. (2014)
20. Shashua, A., Hazan, T.: Non-negative tensor factorization with applications to statistics and computer vision. In: ICML'05, pp. 792–799. (2005)
21. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature, pp. 788–791, vol. 401. (1999)