ICMR '20, October 26–29, 2020, Dublin, Ireland
Proceedings published June 8, 2020

Incorporating Semantic Knowledge for Visual Lifelog Activity Recognition

Min-Huan Fu[1], An-Zi Yen[1], Hen-Hsen Huang[23], Hsin-Hsi Chen[13]
[1]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[2]Department of Computer Science, National Chengchi University, Taipei, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
{mhfu, azyen}@nlg.csie.ntu.edu.tw, hhhuang@nccu.edu.tw, hhchen@ntu.edu.tw

ABSTRACT

The advance in wearable technology has made lifelogging more feasible and more popular. Visual lifelogs collected by wearable cameras capture every single detail of individual's life experience, offering a promising data source for deeper lifestyle analysis and better memory recall assistance. However, building a system for organizing and accessing visual lifelogs is a challenging task due to the semantic gap between visual data and semantic descriptions of life events. In this paper, we introduce semantic knowledge to reduce such a semantic gap for daily activity recognition and lifestyle understanding. We incorporate the semantic knowledge derived from external resources to enrich the training data for the proposed supervised learning model. Experimental results show that incorporating external semantic knowledge is beneficial for improving the performance of recognizing life events.

CCS CONCEPTS

• Computing methodologies → Activity recognition and understanding • Information systems → Personalization

KEYWORDS

Lifelog, Lifelog Activity Recognition, NTCIR Lifelog Dataset, Semantic Knowledge, Word Embedding

ACM Reference format:

Min-Huan Fu, An-Zi Yen, Hen-Hsen Huang, Hsin-Hsi Chen. 2020. Incorporating Semantic Knowledge for Visual Lifelog Activity Recognition. In *2020 International Conference on Multimedia Retrieval (ICMR '20), October 26–29, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3372278.3390700

1 INTRODUCTION

The development of devices for multimedia data capturing, such as video cameras or smartphones, has been long benefiting us for personalized daily life recording. The collected data are generally referred to as lifelogs, most of which are typically stored in digital format nowadays.

Personalized multimedia data captured by lifelogging devices present various aspects of an individual's life experience, offering a rich resource for lifestyle understanding and memory recall. With such a huge amount of data, one could hardly access the desired information without an efficient system for indexing and organizing collected data. Building such a system for lifelogs can, however, be very challenging due to the lack of contextual information as well as the noise in the records.

For supporting memory recall such as searching specific moments in a lifelogger's life or providing summarization on common daily activities, it requires the system to automatically detect and recognize specified activities. The main challenge is that people usually describe their past experiences with textual expressions, while the lifelogs recorded by wearable devices are visual data. It is hard to perform activity recognition without knowing semantic contents present in lifelog images. Thus, introducing external knowledge for semantic content analysis of lifelog images is highly demanded for lifelog system development.

The semantic gap between visual and textual domains poses a serious challenge for multimedia lifelog access. In order to reduce the semantic gap, it is tempting to employ pre-trained computer vision (CV) models to extract semantic contents from lifelog images. Specifically, we could identify the place or the scene of an image with a dedicated classifier for place recognition, or detect multiple objects in an image with a detector trained for common objects recognition. In this paper, we refer to the CV models and their outputs as *visual concept detectors* and *visual concepts*, respectively. A fusion of the visual concepts from different types of concept detectors could give us a shallow semantic interpretation for each lifelog image, but would also suffer from false detections when the image has poor quality.

On the other hand, we observed that there is still a gap between extracted visual concepts and the description of related events. The underlying reason for this semantic gap can be found in the dataset for training visual concept detectors, for most of the annotations in these datasets are rather low-level descriptions such as concrete objects and places. This results in the problem that most of the time we could not find exact matching words between low-level visual concepts and their related event

ICMR '20, October 26–29, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06...$15.00
https://doi.org/10.1145/3372278.3390700

descriptions. For example, the place concept "kitchen" might give us a clue for knowing the image is about "cooking", or the object concept "steering wheel" might imply the semantic activity "driving".

Motivated by the recent success of word embeddings [12] for capturing semantic word relationships in various natural language processing (NLP) tasks, we attempt to incorporate word embeddings into our proposed framework to improve semantic reasoning. We expect the introduction of such semantic knowledge to be beneficial for two aspects of visual lifelog analysis: (1) to reduce the noise caused by false detections of visual concept detectors, and (2) to enhance the relatedness between the visual concepts and the descriptions of their related events.

Following the pilot task proposed by Gurrin et al. [5], we formulate the task toward real-world lifelog use, namely, lifelog activity recognition. The lifelog activity recognition is aimed at automatic recognition of lifelog data in terms of activities of daily living. We discuss the effect of the introducing semantic knowledge for recognizing lifelog activity with experiments.

The contributions of this paper are threefold. (1) We propose a comprehensive system for visual lifelog activity recognition. (2) We introduce word embedding into our proposed framework to reduce the semantic gap between visual and textual domains. (3) The proposed model for lifelog activity recognition incorporates additional textual features without manual annotation and generally improves the performance.

The rest of this paper is organized as follows. Section 2 reviews the related work on lifelog research. Section 3 introduces the dataset adopted in this work, as well as the data pre-processing. Sections 4 describes the methods for lifelog activity recognition. Experimental results are shown and discussed in Section 5. Section 6 concludes the remarks and addresses some future work.

2 RELATED WORK

Lifelogging describes the process to (passively) capture and record personal life experiences of an individual, namely, the lifelogger. As summarized in the study [6], lifelogging could serve to support memory recall or behavior analysis, or for healthcare uses such as lifestyle monitoring, dietary monitoring, etc.

Human activity recognition has been an important topic in computer vision due to its various applications such as content-based video analysis, video surveillance, and human-computer interaction [3]. Numerous research efforts are made to build datasets consisting of annotated short videos for activity recognition, such as UCF101 [17], Sports-1M [9], and ActivityNet [2].

We view visual lifelog activity recognition as a special case of human activity recognition with the egocentric. The egocentric activity recognition has been attracting much attention due to its potential of providing personalized support for various applications. For example, Pirsiavash and Ramanan [14] proposed a dataset consisting of 18 types of activities of daily living, as well as a method for detecting egocentric activities using temporal structure and interactive models of objects. Singh et al. [16] proposed a three-stream convolutional neural network using egocentric cues by capturing hand pose and head motion. However, most previous methods focus on video recognition in short time period, which might not be suitable for long-term lifelog data. Effective approaches to the recognition of lifelog data collected in a longer time span still remain to be explored. Besides, previous works on visual lifelog recognition [1,11] represent visual concepts as the output vectors from pre-trained convolutional neural networks (CNNs). The semantic meaning of each concept is ignored.

In this paper, we refer to the distributed word representation derived from external resources as the *semantic knowledge*. The distributed word representations, also known as word embeddings, are shown to be capable of modeling semantic word relationships. In addition, the learned embeddings also display some linguistic patterns, which can be represented as linear translations in the embedding space.

Moreover, introducing semantic knowledge has been shown beneficial for activity recognition in a low-resource scheme. For example, Jain et al. [8] and Demirel et al. [4] proposed similar methods for zero-shot activity classification based on semantic word embeddings. Zellers and Choi [20] explored large-scale zero-shot activity recognition by modeling the visual and linguistic attributes of action verbs with dictionary definitions and word embeddings.

3 DATASET

We adopt the NTCIR-14 Lifelog Dataset proposed by Gurrin et al. [6], due to the amount and the richness of the contents. The NTCIR-14 Lifelog Dataset contains a collection of multimodal data over 43 days acquired by two active lifeloggers. It consists of 81,474 images, covering realistic topics towards real-world information needs for lifelog applications. We refer to the two lifeloggers as User 1 and User 2, respectively, in the later discussions. Besides, the dataset consists of Multimedia content, Biometrics information, and Human activity data.

3.1 Data Enhancements

In the NTCIR-14 Lifelog Dataset, each wearable camera image is already associated with three types of visual concepts: *placeAttr*, *placeType* and *objectsMS*. The first two visual concepts describe attributes and categories of the place in the image. The remaining one describes the category and location of the detected objects.

Though the visual concept detectors could help for visual understanding, they are pre-trained on datasets with only a small, closed-class set of labels. In order to benefit from larger external resources, we further employ Google Cloud Vision API[1] to extract more visual concepts, including *labelsG* and *objectsG*. The former concept is the visual semantic label, describing the properties about entities across a wide group of categories in the image. The latter concept describes both prominent and less-significant objects within a larger set of object types. At most ten semantic labels and ten objects are extracted, and are associated to each

[1] Google Cloud Vision API. https://cloud.google.com/vision/

lifelog image along with the original concepts. Fig. 1 shows an example of visual concepts.

PlaceAttr: no horizon, man-made, natural light, enclosed area, foliage, open area, wood, leaves, trees, metal
PlaceType: patio, junkyard, tree house, staircase, restaurant patio
ObjectMS: person
LabelsG: barbecue grill, cuisine, barbecue, outdoor grill, kitchen appliance, grilling
ObjectsG: kitchenware, tableware

Figure 1: An example of lifelog image and associated concepts.

3.2 Visual Lifelog Activity Recognition Dataset

Based on previous studies [5,6] and our observation on the dataset, we identify ten common daily activities for both lifeloggers, as listed below:

1. *Traveling*. Travelling or transporting (car, bus, boat, airplane, train, etc.)
2. *Face-to-face interacting*. Face-to-face interaction with people at home or in the workplace (excluding social interactions)
3. *Using a computer*. Using desktop computer, laptop, tablet or smartphone
4. *Cooking*. Preparing meals (include making tea or coffee) at any location
5. *Eating*. Eating meals in any location, excluding moments when drinking alone
6. *Houseworking*. Working in the home (e.g., cleaning, gardening)
7. *Relaxing*. Relaxing at home (e.g., TV, having a drink)
8. *Reading or writing*. reading or writing on any form of paper
9. *Socialising*. Socialising outside the home or office
10. *Shopping*. Shopping in a physical shop (not online)

We manually annotated partial dataset that covers data from both lifeloggers, resulting in a total of 33,463 labels within 15,189 images from User 1 and 16,293 images from User 2 while ignoring images with poor qualities. Data without our manual annotation are neither used for training nor testing the model. The statistics of our annotations are listed in Table 1.

Table1: Statistics of our manual annotations.

User ID	Labeled / Total Images	Number of Labels
User 1	15,189 / 64,073	13,835
User 2	16,293 / 17,599	19,628

We also show the number of labels of each activity class in Table 2, with the ten activities mentioned in the previous section numbered from 01 to 10, sequentially. Note that the large disparity in the number of labels for each class is caused by the difference in duration and frequency between activities.

Table 2: Number of labels in the manual annotation.

	01	02	03	04	05	06	07	08	09	10
U1	2,393	1,324	5,204	405	676	23	1,305	815	1,234	456
U2	305	2338	11,340	485	1,042	371	2,666	652	350	79

4 METHODOLOGY

As multiple activities may take place at the same time, we formulate visual lifelog activity recognition as a multi-label classification problem. Given an image in the collected lifelog data, our goal is to build a model to tell whether a daily activity happened in that image, and which type of the activity was.

4.1 Input Features

Visual Feature. We extract dense feature vectors with the common VGG-19 model [15]. We down-sampled each lifelog image by a factor of 4, and then applied global average pooling to the last convolutional layer, producing a 512-dimensional descriptor for each lifelog image.

Textual Feature. We select four types of the visual concepts mentioned in Section 3.1 – say, *placeAttr*, *placeType*, *labelsG* and *objectsG*, and project them to the embedding space with the pre-trained embedding GloVe [13] to include semantic information of each concept. By doing so, we could obtain an unordered set of visual concepts associated to each lifelog image. The standard 300-dimensional GloVe, pre-trained on the 6B tokens, is adopted in this work. By employing pre-trained word embeddings, the model is enabled to leverage the external semantic knowledge derived from huge volume of real-world corpora.

4.2 Model Structure

The proposed DNN models take the features of lifelog images in both visual and textual modalities as input, as shown in Fig. 2 (b). Note that the textual features of visual concepts are represented as unordered sets of vectors, which means that the model should be order-independent to its input, and thus common neural network structures for ordered texts, e.g., convolutional neural networks (CNNs) or recurrent neural networks (RNNs), are hardly applicable in this case.

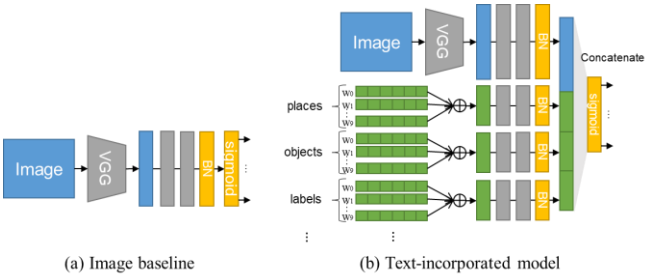

Figure 2: The structure of proposed DNN models.

Inspired by the simple but effective deep averaging networks on text classification tasks [7], we adopt similar composition functions to deal with unordered input. For a set of visual concepts X, we obtain the set representation $\boldsymbol{z}_t$ by aggregating embedding

vectors $\boldsymbol{v}_w$ of $w \in X$ with an unordered composition function g. A simple choice of g can be an averaging operator:

$$\boldsymbol{z}_t = g(w \in X) = \frac{1}{|X|}\sum_{w \in X} \boldsymbol{v}_w \quad (1)$$

We transform $\boldsymbol{z}_t$ with two fully-connected layers, followed by the batch normalization as a regularizer. Note that we obtain the above representation for each type of visual concept, so the textual representation obtained by the DNN model is actually a concatenation of transformed $\boldsymbol{z}_t$'s from different types of visual concepts.

We also transform visual feature vector $\boldsymbol{z}_v$ by two fully-connected layers and the batch normalization layer. Features from the two modalities are combined with vector concatenation, and passed to a sigmoid output layer for multi-label classification. The dimension of the output layer is aligned to the number of activity types. A naïve model that exploits only visual feature vector $\boldsymbol{z}_v$ serves as our baseline model, as shown in Fig. 2 (a).

4.3 Weighted Concept Aggregation

The method in Section 4.2 gives equal importance to each visual concept word. However, as argued in Tang et al. [19], combining outputs of different visual concept detectors may give us more information about the image. There are also risks to include noise into the set of visual concepts due to the false detections. For estimating the importance of each word, we integrate a word relatedness matrix of visual concept words into our DNN model. The semantic relatedness or similarity between words is commonly captured as the cosine similarity in the embedding space [12]. Instead of simple inner products, we adopt a bilinear form of vectors to include a trainable matrix $\mathbf{B}$ shared among all pairs of concept words. The normalized relatedness between visual concepts w and w' can be written as:

$$rel(w, w') = \frac{1}{|\boldsymbol{v}_w||\boldsymbol{v}_{w'}|}\boldsymbol{v}_w^{\mathrm{T}}\mathbf{B}\boldsymbol{v}_{w'} \quad (2)$$

For each concept set X, the expanded concept set X' is formed by adding $k - |X|$ "empty" word ε with $\boldsymbol{v}_\varepsilon = \mathbf{0}$, where k is the maximum number of concept words and d is the dimension of word embeddings. Then, the row vectors $\boldsymbol{v}_w^{\mathrm{T}}$ of the concept words $w \in X'$ are vertically stacked into a $k \times d$ semantic matrix $\mathbf{M}$, and the relatedness matrix $\mathbf{R}$ can be calculated as $\mathbf{R} = \mathbf{M}\mathbf{B}\mathbf{M}^{\mathrm{T}}$, of which each entry serves as relatedness between pairs of visual concepts.

We may interpret each row in $\mathbf{R}$ as "how much each concept word is supported by other words," and we expect the model to give more weights to those visual concepts that accumulate higher correlations with other concepts in the same image. In this sense, the sum over each row of the relatedness matrix $\mathbf{R}$ is collected to derive weighting vector $\boldsymbol{a}^{\mathrm{T}} = [\sum_{j=1}^{k}\mathbf{R}_{1,j} \cdots \sum_{j=1}^{k}\mathbf{R}_{k,j}]$ for vector aggregation, where the i-th entry $\boldsymbol{a}_i$ is the weighting factor of the i-th word $\boldsymbol{v}_{w_i}$ in X'. The representation $\boldsymbol{z}_t'$ is then derived by the weighted composition function h, conditioned on $\boldsymbol{a}$:

$$\boldsymbol{z}_t' = h(w \in X'; \boldsymbol{a}) = \frac{1}{\sum \boldsymbol{a}_i}\left(\boldsymbol{a}_1\boldsymbol{v}_{w_1} + \cdots + \boldsymbol{a}_k\boldsymbol{v}_{w_k}\right) \quad (4)$$

where the plus sign denotes the vector addition. Note that the weighted composition function h actually computes the column sum of the semantic matrix $\mathbf{M}$ weighted by vector $\boldsymbol{a}$. We may rewrite h in a matrix multiplication form:

$$\begin{aligned}\boldsymbol{z}_t'^{\mathrm{T}} &= h(\mathbf{M}; \boldsymbol{a}) = \frac{1}{\sum \boldsymbol{a}_i}\boldsymbol{a}^{\mathrm{T}}\mathbf{M} \\ &= \frac{1}{\sum \boldsymbol{a}_i}\left[\sum_{i=1}^{k}\boldsymbol{a}_i\mathbf{M}_{i,1} \cdots \sum_{i=1}^{k}\boldsymbol{a}_i\mathbf{M}_{i,d}\right] \\ &= \frac{1}{\sum \boldsymbol{a}_i}\left(\boldsymbol{a}_1\mathbf{M}_{1,*} + \cdots + \boldsymbol{a}_k\mathbf{M}_{k,*}\right)\end{aligned} \quad (3)$$

where $\mathbf{M}_{i,*}$ denotes the i-th row in $\mathbf{M}$. By substituting $\boldsymbol{v}_{w_i}$ with $\mathbf{M}_{i,*}^{\mathrm{T}}$ in the rightmost part of (4), we have $\boldsymbol{z}_t' = \frac{1}{\sum \boldsymbol{a}_i}\left(\boldsymbol{a}_1\mathbf{M}_{1,*}^{\mathrm{T}} + \cdots + \boldsymbol{a}_k\mathbf{M}_{k,*}^{\mathrm{T}}\right)$, which turns out to be equivalent to (5). By rewriting the formula, the weighted composition can be combined to the neural network with a simple matrix multiplication.

The aggregated concept representation $\boldsymbol{z}_t'$ is also transformed through two fully-connected layers followed by batch normalization, and concatenated with visual feature $\boldsymbol{z}_v$ before the sigmoid output layer. The whole structure is shown in Fig. 3.

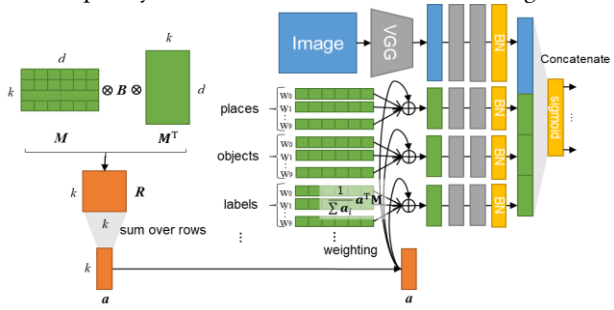

Figure 3: The DNN model using weighted aggregation.

Alternatively, we may also exploit the relation between visual concepts and the descriptions of activities listed in Section 3.4. For example, the concept *food* is considered highly related to the description "*eating meals in any location* ...," due to the high similarities between *food* and *eating*, and *food* and *meals*. For description D of each daily activity, we add empty words ε to align the length of D to the maximum length of description, and encode D into a sequence of word embeddings $\mathrm{V}_D = \{\boldsymbol{v}_D^1, \boldsymbol{v}_D^2 \cdots \boldsymbol{v}_D^{|\mathrm{V}_D|}\}$. The normalized relatedness of each visual concept w and description D is written as:

$$rel(w, D) = \max_{\boldsymbol{v}_D \in \mathrm{V}_D} \frac{1}{|\boldsymbol{v}_w||\boldsymbol{v}_D|}\boldsymbol{v}_w^{\mathrm{T}}\mathbf{B}\boldsymbol{v}_D \quad (4)$$

in which the most similar word to w in description D is selected.

The relatedness matrix $\mathbf{S}$ can be calculated as matrix multiplication with stacked representations of k concept words and l activity descriptions, as shown in Fig. 4. This results in a $k \times l$ concept word-description relation matrix $\mathbf{S}$. We compute the dot product of $\mathbf{S}^{\mathrm{T}}$ and the semantic matrix $\mathbf{M}$ to obtain l aggregations of visual concepts weighted by the relatedness to each activity, transform the weighted aggregations with two fully-connected layers and the batch normalization, and concatenate them to produce the final representation $\boldsymbol{z}_t''$.

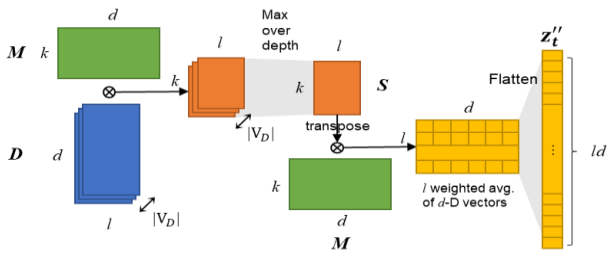

Figure 4: The alternative weighted aggregation for the DNN model.

5 EXPERIMENTS

Experimental results are shown in Sections 5.1. Section 5.2 analyzes the performances on different sizes of the training data. Section 5.3 analyzes the performances of the model with different concept types. Section 5.4 analyzes performance of training and testing on different lifeloggers.

5.1 Experimental Results

We adopt the rectified linear unit (ReLU) as the activation function of the latent layers, and the sigmoid function at output layer for multi-label classification. The dimension of the latent layers is set to 128 and 75 for visual and textual features, respectively. We list the variations of models in Table 3. The visual feature of lifelog images extracted by the VGG-19 model is used in all the above models. For comparison, we also report the performance of the model based on only textual features (*TextOnly*).

Table 3: Variations of the DNN models.

Model	Description
ImageOnly	Model using only visual features (Fig. 2 (a))
TextOnly	Model using only textual features
TextAvg	Model using visual features and averaged textual features (Fig. 2 (b))
TextCorr	Model using visual features and averaged textual features, weighted by the correlation between visual concepts (Fig. 3)
TextSim	Model using visual features and averaged textual features, weighted by the relation between concepts and activity description (Fig. 4)

The performance of different variations of the proposed DNN models is shown in Table 4. All the reported performance scores are the average of 5-fold cross validation. F-score (F1) is the main metric for performance evaluation. The micro-F1 score is computed as the harmonic mean of precision (P) and recall (R). The macro-F1 score is computed separately and averaged with even weights.

Since the dataset contains about 10% positive labels, the expected value of the precision score for random guess is about 10%. As can be seen, the DNN models using both visual and textual features generally outperform those using only unimodal feature. The *textSim* model and *textCorr* model improve the F1-scores the most, reaching 70.96% micro-F1 score and 52.1% macro-F1 score, respectively. The result implies that textual features from the visual concepts can provide complementary information to visual features and generally improve model performance. To our surprise, the *textOnly* model performs on a par with the *imageOnly* model, implying that the visual concepts might contain as much information as the visual features, while we may expect some information loss during the extraction of visual concepts.

We also observe that *textCorr* and *textSim* do not improve much on micro-F1 score in comparison with *TextAvg*, but achieve higher macro-F1 scores. This result implies that the models with weighted composition functions have better performance on recognizing rare activities. On the other hand, the weighted composition functions help improve the recall scores of the model, but slightly decrease the precision score: *textCorr* achieves the best recall score of 66.18%, while *textAvg* achieves the best precision score of 80.55%. The result suggests that we may choose different composition functions so that the model will favor precision or recall for different purposes of use.

Table 4: Performance of different variations of models.

		P	R	Micro-F1	Macro-F1
Random		10.53%	50.21%	17.39%	13.92%
Uni-modal	ImageOnly	76.73%	59.21%	66.63%	41.60%
	TextOnly	77.56%	59.44%	67.28%	44.45%
Bi-modal	TextAvg	**80.55%**	62.65%	70.42%	49.72%
	TextCorr	75.90%	**66.18%**	70.65%	**52.10%**
	TextSim	78.99%	64.55%	**70.96%**	51.23%

We also report the label-wise performance of different models in Table 5 and Table 6. The best improvement by incorporating textual features can be seen in the activities *cooking* and *shopping*, with an increase on the F1-score of about 15%.

Table 5: Label-wise performance of different DNN models.

	ImageOnly			TextAvg		
	P	R	F1	P	R	F1
Travel	89.39%	65.11%	73.67%	96.24%	79.05%	86.01%
Interaction	50.04%	23.61%	31.20%	61.71%	32.97%	42.22%
Computer	84.47%	84.03%	84.17%	87.64%	84.31%	85.83%
Cooking	51.65%	28.67%	35.86%	65.55%	42.14%	50.72%
Eating	62.05%	31.03%	40.54%	63.89%	37.40%	46.13%
Housework	43.32%	12.05%	18.34%	46.43%	16.91%	24.36%
Relaxing	50.65%	36.03%	40.91%	51.68%	37.29%	42.67%
Reading	46.65%	11.31%	15.78%	48.70%	13.19%	18.93%
Socializing	42.49%	15.78%	22.53%	56.94%	23.52%	32.20%
Shopping	62.86%	55.18%	53.02%	76.25%	68.53%	68.09%

Table 6: Label-wise performance of different composition function.

	TextAvg			TextCorr		
	P	R	F1	P	R	F1
Travel	96.24%	79.05%	86.01%	87.37%	86.38%	86.57%
Interaction	61.71%	32.97%	42.22%	51.86%	43.05%	46.30%
Computer	87.64%	84.31%	85.83%	85.64%	85.27%	85.28%
Cooking	65.55%	42.14%	50.72%	63.66%	45.83%	52.06%
Eating	63.89%	37.40%	46.13%	63.45%	39.32%	47.47%
Housework	46.43%	16.91%	24.36%	45.41%	18.92%	25.23%
Relaxing	51.68%	37.29%	42.67%	50.10%	45.08%	46.66%
Reading	48.70%	13.19%	18.93%	50.14%	18.89%	24.19%
Socializing	56.94%	23.52%	32.20%	50.20%	27.74%	35.29%
Shopping	76.25%	68.53%	68.09%	77.95%	71.56%	71.90%

It is also worth noting that both activities have rather few positive samples in the dataset, less than 3%. On the other hand,

the weighted composition function improves the recall score of all activities, particularly in the activity *face-to-face interacting* with an increase of about 10%. However, the precision score generally decreases, so the improvement of F1-score is not as significant as in Table 5.

5.2 Performance vs. Dataset Size

In this section, we report the performance and the difference of performance on different sizes of the training data in Fig. 5. We randomly sample partial training data to obtain smaller training set, and keep the left-out testing data unchanged.

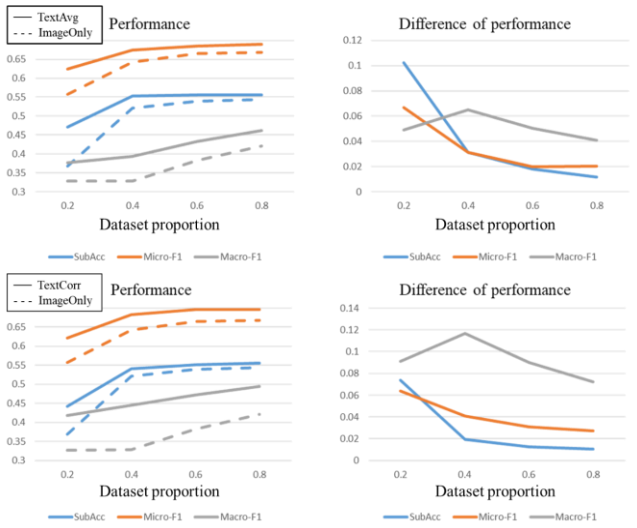

Figure 5: Performances on different sizes of the training data.

The dashed line and the solid line in the left plot in Fig. 5 show the performance of *imageOnly* and *textAvg* (*textCorr*), respectively. The right plot in Fig. 5 shows the increase of performance in *textAvg* (*textCorr*) model. As can be seen, *textAvg* (*textCorr*) generally outperforms *imageOnly* with different sizes of the training data, and particularly boost the performance with smaller training data.

5.3 Performance vs. Concept Types

To test the importance of each type of concept, we train different DNN models, each of which includes only single type of concept aggregated with the weighted averaging function (*textCorr*). The results are shown Table 7.

Table 7: Performance vs. different types of visual concepts.

	Precision	Recall	Micro-F1	Macro-F1
ImageOnly	76.73%	59.21%	66.63%	41.60%
+ placeAttr	77.89%	63.27%	69.68%	49.35%
+ placeType	77.72%	61.08%	68.30%	47.62%
+ labelsG	77.20%	62.12%	68.72%	47.30%
+ objectsG	78.49%	61.74%	68.95%	45.99%

In Table 7, the model using only the *placeAttr* concepts reaches overall highest scores in most of the metrics. As for the macro-F1 score, the *placeAttr* feature achieves an improvement of about 8%, which implies that knowing the attribute of the place is particularly crucial for recognizing rare activities in the dataset. Regardless of which type of visual concept is used, the model incorporated with textual features generally outperforms the model that uses only visual features in our experiments.

5.4 Results on Different Users

Since different lifeloggers might have quite different lifestyles, the collected images of the same activity from different lifeloggers are not always visually similar. That is, we expect the model detects less correctly when training data and test data are collected by different lifeloggers. To test the adaptability of the model across lifeloggers, we train the model on User 1's data and test on User 2's data, and vice versa. The results are shown in Table 8, in which the performance has significantly dropped compared with the scores reported in Table 4. Interestingly, the *textOnly* model achieves better performance than the *ImageOnly* model in both cases, implying that textual features are more consistent than visual features across different lifeloggers.

Table 8: performance of training and testing on different lifeloggers.

		P	R	Micro-F1	Macro-F1
U1 →U2	ImageOnly	81.90%	41.03%	54.67%	22.86%
	TextOnly	84.48%	43.85%	57.73%	32.04%
U2 → U1	ImageOnly	53.85%	39.93%	45.86%	19.95%
	TextOnly	60.19%	47.87%	53.33%	29.62%

6 CONCLUSION AND FUTURE WORK

Automatic detection and recognition of the lifelog activities plays an important role for understanding the patterns of daily living. In this paper, we attempt to introduce the semantic knowledge into lifelog systems. For visual lifelog activity recognition, we incorporate the textual features of the visual concepts aggregated in an unordered fashion to enrich the training data for supervised DNN models. Experimental results show that our purposed method for incorporating semantic knowledge is effective for improving the performance of lifelog systems, compared with the baseline models.

Our future work is to incorporate knowledge base resources such as ConceptNet [18] to support semantic reasoning by modeling the commonsense relationships between words, or introduce visual-grounded embeddings to encode textual data [10]. In this way, the system can capture the relationships between visually similar words. Moreover, other metadata or contextual information, such as the biometrics and GPS information, can also be incorporated as a part of the multimodal representation.

ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-109-2634-F-002-040-, MOST-109-2634-F-002-034-, MOST-108-2218-E-009-051-, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

REFERENCES

[1] Fatma Ben Abdallah, Ghada Feki, Mohamed Ezzarka, Anis Ben Ammar, and Chokri Ben Amar. 2018. Regim Lab Team at ImageCLEF Lifelog Moment Retrieval Task 2018. In *CLEF* (Working Notes).

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–970.

[3] Guangchun Cheng, Yiwen Wan, Abdullah N. Saudagar, Kamesh Namuduri, and Bill P. Buckles. 2015. Advances in human action recognition: A survey. *arXiv preprint arXiv:1501.05964*.

[4] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. 2017. Attributes2classname: A discriminative model for attribute-based unsupervised zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 1232–1241.

[5] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, Dang Nguyen, and Duc Tien. 2017. Overview of NTCIR-13 Lifelog-2 task. In *NTCIR*.

[6] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Van-Tu Ninh, Tu-Khiem Le, Rami Albatal, Duc Tien Dang Nguyen and Graham Healy. 2019. Overview of NTCIR-14 Lifelog-3 task. In *NTCIR*.

[7] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), volume 1, 1681–1691.

[8] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. 2015. Objects2action: Classifying and localizing actions without any video example. In *Proceedings of the IEEE international conference on computer vision*, 4588–4596.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.

[10] Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative Language Understanding: Large-Scale Visual Grounding with Image Search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 922–933, Melbourne, Australia, July. Association for Computational Linguistics.

[11] Jie Lin, Ana Garcia del Molino, Qianli Xu, Fen Fang, Vigneshwaran Subbaraju, Joo-Hwee Lim, Liyuan Li, and Vijay Chandrasekhar. 2017. VCI2R at the NTCIR-13 Lifelog-2 Lifelog Semantic Access Task.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP), 1532–1543.

[14] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2847–2854. IEEE.

[15] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[16] Suriya Singh, Chetan Arora, and C. V. Jawahar. 2016. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2620–2628.

[17] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

[18] Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[19] Tsun-Hsien Tang, Min-Huan Fu, Hen-Hsen Huang, Kuan-Ta Chen, and Hsin-Hsi Chen. 2018. Visual Concept Selection with Textual Knowledge for Understanding Activities of Daily Living and Life Moment Retrieval. In *CLEF* (Working Notes).

[20] Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. *arXiv preprint arXiv:1707.09468.Conference* Name:ACM Woodstock conference