Incorporating External Textual Knowledge for Life Event Recognition and Retrieval

Min-Huan Fu[1], Chia-Chun Chang[1], Hen-Hsen Huang[2,3] and Hsin-Hsi Chen[1,3]

[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, Taipei, Taiwan
{mhfu, ccchang}@nlg.csie.ntu.edu.tw;
hhhuang@nccu.edu.tw; hhchen@ntu.edu.tw

Abstract. This paper presents our approach to the task of NTCIR-14 Lifelog-3. We participated in two of the subtasks, lifelog semantic access task (LSAT) and lifelog activity detection task (LADT). We attempt to reduce the semantic gap presented in lifelog tasks by introducing textual knowledge derived from external resources. In both subtasks, we extract additional visual concepts with computer vision models, and then incorporate both official and additional concepts into our system using pre-trained word embeddings, in which textual knowledge is inherent. For LSAT, we propose an interactive system that automatically suggests users a list of candidate query words, and adopt probabilistic relevance-based ranking function for retrieval. Our system also allows users to refine the retrieval results by filtering irrelevant images out. For LADT, we incorporate visual concepts into our supervised learning framework. We first encode visual concepts with pre-trained word embeddings, and perform unordered aggregation to produce order-independent representation of visual concepts. In terms of performance, our systems achieve 0.4727 of mean average precision in LSAT and 0.5439 of F1 score in LADT.

Keywords: Lifelog, Interactive System, Word Embedding, Multimodal Learning

Team Name: nlg301

Subtasks:
Lifelog Semantic Access Task (LSAT) (English),
Lifelog Activity Detection Task (LADT) (English)

1 Introduction

The increasing availability of dedicated lifelogging devices has been offering a novel choice for daily life recording. Personalized data captured by lifelogging devices offers a rich resource for daily life understanding and memory recall. Typically, these devices capture life events in a passive fashion and generate a huge volume of multimedia archives as time proceeds. Efficient approaches for users to organize and access collected lifelog data are thus highly demanded.

In NTCIR-14 Lifelog [1], Lifelog Activity Detection Task (LADT) and Lifelog Semantic Access Task (LSAT) are aimed at addressing the problems of indexing and retrieving lifelog data. The lifelog data consists of multimodal information (images, body metrics, semantic place/activity information), and is enriched with semantic visual concepts extracted with computer vision (CV) models. Extended from our previous work [2], this paper explores the potential for applying natural language processing (NLP) techniques to the lifelog task. Specifically, we exploit external textual information by introducing the semantic word embeddings, to reduce semantic gap between textual queries and visual concepts for LSAT task and to enrich training data of supervised learning for LADT task.

2 Related Work

One of the major challenges for accessing multimedia lifelog is the so-called semantic gap between the visual information from lifelog images and textual information from event-based queries [2,3]. As a common approach to this issue, existing work introduces external resources such as ImageNet [4] or Place365 [5] to bridge the semantic gap. Knowledge derived from these resources are either incorporated as dense feature vectors [3,6,7,8], or further expanded with external textual knowledge including semantic word embeddings or WordNet ontology [2,9,10].

Motivated by the remarkable progress of deep neural networks in computer vision and natural language processing, recent lifelog researches mainly focus on deep learning-based approaches. For example, previous work [3] proposed a retrieval scheme based on feature importance of deep features learned with conditional random field (CRF). Another work [6] proposed a deep learning strategy for lifelog event detection, exploiting a fusion of multimodal features. On the other hand, the work [9] incorporates external textual knowledge for visual indexing and query expansion with multimodal attentive LSTMs and semantic word embeddings, which is similar to our work. We present this work as an extension of our previous work [2], in which we introduced external textual knowledge to our retrieval system using pre-trained word embeddings.

3 Data Preprocessing

3.1 Visual Concept Indexing

The NTCIR-14 Lifelog organizers provide a set of visual semantic concepts extracted with concept detectors, including image attributes, places and objects [1]. In addition to official concepts, we further extract more visual concepts with Google Cloud Vision API[1]. At most ten visual semantic labels and at most ten objects are detected with the label detector and object detector, respectively. The extracted concepts are associated to each lifelog image along with the official concepts.

3.2 Image Preprocessing

We observed that advanced CV models are prone to error for images with poor quality. To ensure the quality of input images to the models, we apply lens calibration, followed by blurriness and color diversity detection [2].

Lens Calibration. The images in the dataset are captured by the OMG Autographer[2], of which the camera is a 136° wide-angle lens. This results in slight distortion of captured images, leading erroneous outputs of the CV models. For example, Lens calibration is performed on all images with commercially available photo editing software. We show some before and after results of calibration in Fig. 1. The upper part shows an example corrected by the calibration, and the lower part shows that the calibration has little effect on correctly tagged images.

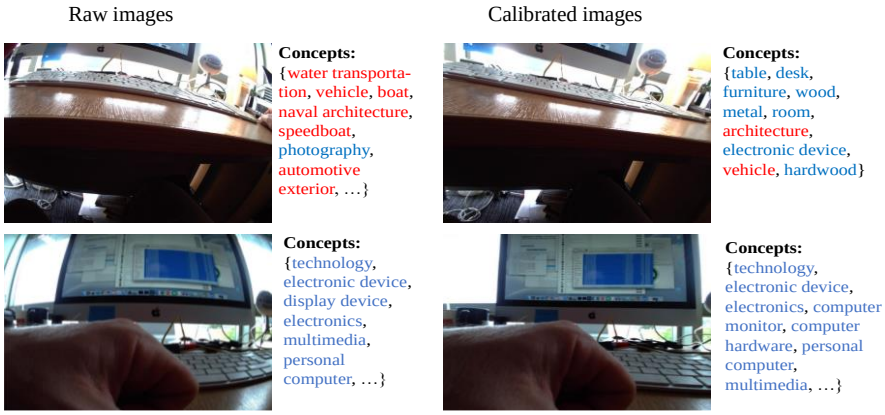


Fig. 1. The effect of lens calibration on Google Cloud Vision API. Concepts shown in blue are considered relevant, while concepts in red are irrelevant.

[1] https://cloud.google.com/vision/
[2] https://vandrico.com/wearables/device/omg-autographer/

Image Filtering. We prune low quality images with blurriness and color diversity detection. The blurriness metric is defined based on the variation of the Laplacian. Each image is convolved with a 3×3 Laplacian filter, and the blurriness measurement is calculated as the variance of the convolved result. Images with low variance of are considered blurry and undesirable. Besides, images with high color homogeneity are also considered uninformative, and can be detected with quantized color histograms.

4 LSAT

4.1 Retrieval Framework

In our retrieval framework, lifelog images are represented as short documents consisting of concept words, and are associated with the metadata recorded from lifelogging device, e.g., time information. Given a set of query words, we apply BM25 [11] as the ranking function for document retrieval. BM25 measures the probabilistic relevance between two sets of concept words based on term frequency (TF) and inverse document frequency (IDF). In this way, rare concepts are given more importance and are more likely to be captured in the retrieval scheme. The retrieval results are sorted by descending order of ranking scores. We describe the query formulation process of our system in the following section.

4.2 Interactive Retrieval System

This section introduces our interactive retrieval system and shows how it benefits the LSAT subtask. The main purpose of this system is to allow users to refine the retrieval results in an interactive fashion. The system is built as a search engine that provides a customized retrieval operation for users. Previous work [3,6] considers that crucial components for lifelog understanding include what, where, and when the lifelogger does. The components can be covered by visual concepts of lifelog images along with the time information recorded by lifelogging devices. Accordingly, we provide three options in our system allowing users to specify the queries and the search constraints, but do not provide more options because the time of retrieval is limited.

The retrieval process of our interactive system includes the following steps: choosing query words, deciding time interval and user ID, and selecting relevant images. The system automatically suggests the user with a list of semantically related concept words to each query word, as shown in Fig. 2. The users can choose the suggested query words, or manually input words as additional query terms. Finally, the API will return all images that match the given query words and constraints.

Query Suggestion. Our query expansion strategy is similar to a previous work [9], in which the retrieval systems offer the suggested concept words related to each query. The main difference is that we restrict the search space of the nearest neighbor search to a very small range, including only the extracted concept words presented in the da-

taset. Due to the small vocabulary size (1,907) of extracted concept words, it is computationally feasible to search the nearest neighbors exhaustively in the embedding space. We exploit pre-trained word embeddings to obtain the top K (default K=10) similar concept words by comparing the semantic similarity between concept words and query terms. The results are used as the query word suggestion of our system.

Refining Retrieval Results. The system allows users to remove irrelevant images by clicking the images. We refer to this process as the user refinement on retrieval results. The remaining images are considered relevant and submitted as the final result.

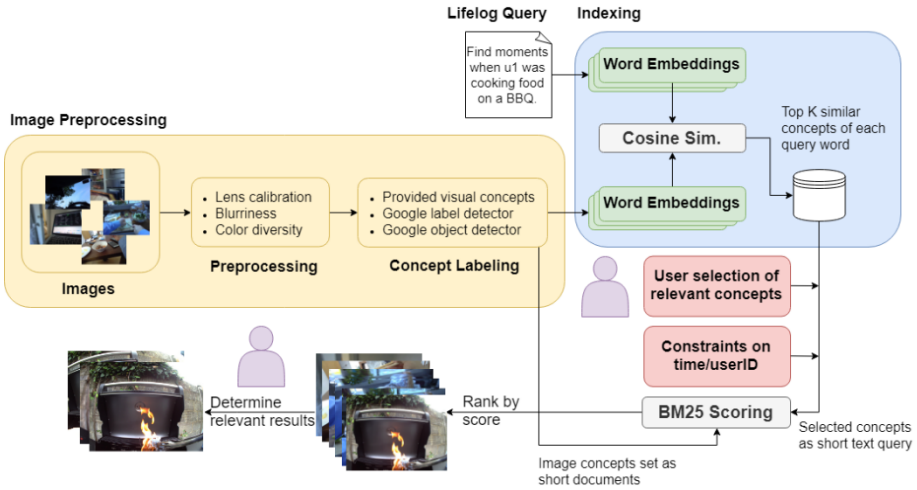


Fig. 2. The framework of our interactive system.

4.3 Official Result

We submitted a total of three runs to the LSAT subtask, including two interactive runs. In the first run, we use all suggested query words for retrieval; this serves as our baseline method. The second and the last runs share the same queries formulated by the same user, while the second run does not include the user refinement process.

The official result is shown in Table 1. We report mean average precision (mAP), P@10, total number of relevant documents retrieved (RelRet), and mean reciprocal rank (MRR) of each submitted run. We also report the average interaction time of the last run. The result shows that our interactive approach for lifelog retrieval substantially improves the overall performance of the LSAT subtask, especially the precision score. It is worth noting that the total number of relevant documents retrieved has slight decrease after performing the user refinement. This may result from the fact that the user of our system is not the lifelogger, and possibly make wrong deletions of the relevant retrieval results.

6

Table 1. Official result of LSAT. Highest score in each column is shown in bold.

Run ID	mAP	P@10	RelRet	MRR	Inter. Time
Run01 (Automatic)	0.0632	0.2375	293	0.3963	-
Run02 (Interactive)	0.1108	0.3750	**464**	0.6571	-
Run03 (Interactive)	**0.1657**	**0.6833**	407	**0.8625**	159.5 s

5 LADT

5.1 Proposed approach

The LADT subtask is aimed at automatically annotating lifelog images with sixteen pre-defined labels of daily activities. As multiple activities may happen simultaneously, the task is regarded as a multi-label classification, which is typically addressed with supervised machine learning approaches. To this end, two of the authors manually annotated the images in four days of each lifelogger as training data. We chose these days for covering most of the activities in the dataset, and the statistics of annotations are reported in Table 2. We take a similar scheme of supervised learning as in the work [6], in which deep neural network (DNN) models are adopted.

Table 2. Number of labels in the training data. The sixteen classes are numbered sequentially: {traveling, f2f interaction, using a computer, cooking, eating, time with children, housework-ing, relaxing, reading, socializing, praying, shopping, gaming, physical activities, creative activities, other activities}

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
u1	711	198	1816	120	137	0	12	429	86	268	0	160	0	64	34	0
u2	78	979	2654	131	356	0	124	827	204	159	0	44	91	0	51	31

5.2 Model Input

Visual Features. Pre-trained convolutional neural network (CNN) models have been shown beneficial for various learning tasks as generic image feature extractors. In this sense, we extract low-level image descriptors with the common VGG-19 model [12], pre-trained on ImageNet1K dataset. We take the global average-pooled output of the layer before *softmax* output layer, resulting in a 512-D feature vector describing each lifelog image.

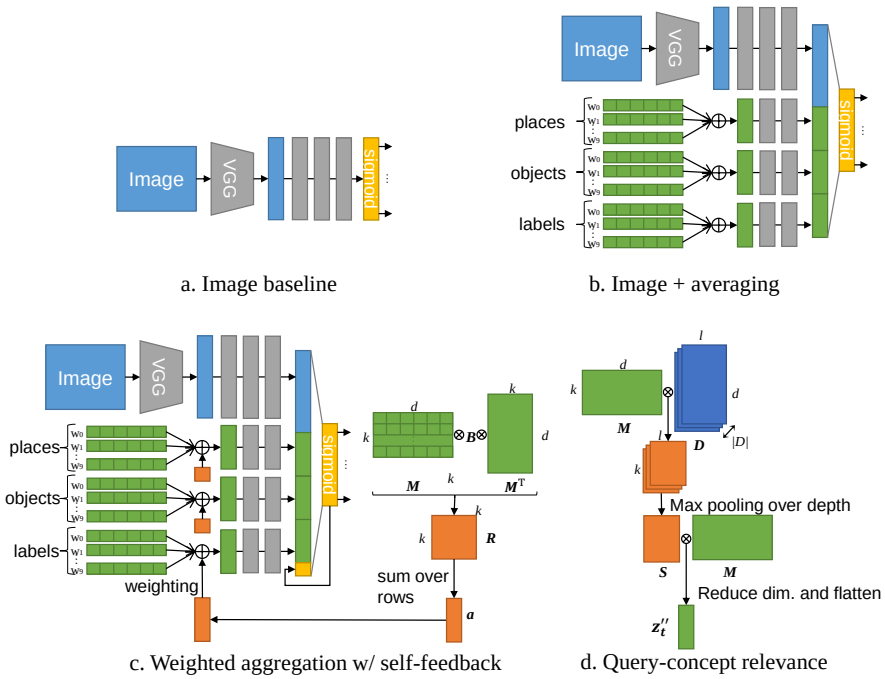
Textual Features (Visual Concepts). Previous work [3,6,7,8] represents high-level visual concepts with the output of deep CNNs, in which each concept is associated with a probability value. While taking fixed-size dense vector as model input is rather convenient, it does not include the semantic meaning of each concept. To represent the semantic information of high-level concepts detected by CV models, we encode each

concept word with GloVe, trained on 6B tokens [13]. By using pre-trained word embeddings, the model is enabled to leverage external textual knowledge derived from huge volume of real-world corpora. We explore various approaches for aggregating set of word embeddings in Section 5.3.

5.3 The DNN Model

The proposed DNN models take features of visual and textual modalities as input, as shown in Fig. 3. One of the main challenges to include set of vectors as neural network input is that common network structures for ordered text, such as CNNs or RNNs, are not applicable in this case. Inspired by the simple but effective deep averaging networks on text classification tasks [14], we adopt similar composition functions to deal with unordered input. For a set of concept words X, we obtain the set representation $\boldsymbol{z}$ by aggregating embeddings $\boldsymbol{v}_w$ of $w \in X$ with the unordered composition function g. A simple choice of g is an averaging operation, as in (1). We further encode $\boldsymbol{z}_t$ with two fully-connected layers.

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



a. Image baseline

b. Image + averaging

c. Weighted aggregation w/ self-feedback

d. Query-concept relevance

Fig. 3. Variations of proposed DNN model.

As mentioned in Section 3.1, concepts extracted from CV models include image attributes, places, objects and semantic labels. We obtain the above representations for

each type of concepts, so the textual features for our DNN model is actually a concatenation of these representations.

We also transform the low-level visual feature vector $\boldsymbol{z_v}$ by fully-connected layers, followed by a dropout layer of rate 0.5 to prevent overfitting. Features from the two modalities are combined with vector concatenation, and passed to a sigmoid output layer for multi-label classification task. The DNN structure is shown in Fig. 3 (b). For practical reason, we exclude activities without any training instance in our model, so the dimension of output layer is 14. A naïve model that exploits only $\boldsymbol{z_v}$ as input serves as our baseline model, as shown in Fig. 3 (a).

5.4 Weighted Concept Aggregation

The method in Section 5.3 gives equal importance to each concept word. However, as argued in [2], combining outputs from CV models may result in redundancy and noise in the set of concepts due to the false positives. Motivated by the concept selection process proposed in [2], we integrate a word relatedness matrix of concept words into our DNN model for estimating the importance of each word. The semantic relatedness between words is commonly captured as the cosine similarity in the semantic vector space [15]. Instead of simple inner product, we adopt a bilinear form of vectors to include a trainable matrix **B** shared among pairs of concept words. The normalized relatedness s of embeddings $\boldsymbol{v_w}$ and $\boldsymbol{v_{w'}}$:

$$s_{w,w'} = \frac{1}{|\boldsymbol{v_w}||\boldsymbol{v_{w'}}|} \boldsymbol{v_w^T} \mathbf{B} \boldsymbol{v_{w'}}. \tag{2}$$

For any set of concept words X, embeddings $\boldsymbol{v_w^{\mathrm{T}}}$ of concept word in X are horizontally stacked into a $k \times d$ semantic matrix **M**, where k is the maximum number of concept words and d is the dimension of word embeddings. For set X with insufficient number of words, **M** is padded to agree the maximum matrix size. The relatedness matrix **R** can be then calculated as $\mathbf{R} = \mathbf{MBM^T}$, of which each element serves as relatedness between concept words.

We may interpret each row in **R** as "how much each concept word is supported by other words," and concept words with more support from others are considered important. In this sense, the sum over each row of **R** is collected to derive the weighting vector $\boldsymbol{a}$ for concept aggregation. The representation $\boldsymbol{z_t'}$ for set of concept words is

$$\boldsymbol{z_t'} = h(\mathbf{M}; \boldsymbol{a}) = \boldsymbol{a^T}\mathbf{M}. \tag{3}$$

The representation is transformed through two fully-connected layers and concatenated with visual feature $\boldsymbol{z_v}$ before the sigmoid output layer. The model also accepts its prediction of images in previous K (default K=3) time steps as an additional input, as the previous images may sometimes be more informative because of the sequential characteristic of lifelog images. We refer this mechanism as a self-feedback approach. The whole structure is illustrated in Fig. 3 (c).

Alternatively, we may also exploit the relation between visual concepts and the descriptions of activities. 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*, *meals*. To reduce the abstractness of the official descriptions, we manually rewrite the descriptions without changing their meanings, and encode them into sequence of embeddings d. The normalized relatedness of each concept word w and description D is written as (4).

$$s_{w,d} = \max_{\boldsymbol{v_D} \in \boldsymbol{D}} \frac{1}{|\boldsymbol{v_w}||\boldsymbol{v_D}|} \boldsymbol{v_w^T} \mathbf{B} \boldsymbol{v_D}. \quad (4)$$

The relatedness matrix can be calculated as matrix multiplication with stacked representations of k concept words and l activity descriptions, as shown in Fig. 3 (d). This results in a $k \times l$ concept word-description relation matrix $\mathbf{S}$. We perform inner product of $\mathbf{S^T}$ and the $\mathbf{M}$ to obtain l aggregations of concepts weighted by the relatedness to each activity. We reduce the dimension of weighted aggregations with fully-connected layers and concatenate them to produce the final representation $\boldsymbol{z_t''}$. (Note that $\boldsymbol{z_t}$, $\boldsymbol{z_t''}$, $\boldsymbol{z_t''}$ are order-independent due to the unordered aggregation strategies.

5.5 Post-processing

We attempt to improve the annotation result by ad-hoc post-processing strategies in the last two runs. In the post-processing stage, we intentionally set prediction scores of the activities traveling, socializing, and shopping to 0 if the image is captured at home or at work. In addition, for image without any annotation, we search its neighboring images in a small context window for agreeing annotations (if any) as its annotation.

5.6 Official Results

We submitted a total of ten runs to the LADT subtask with different model variation, including different strategies to incorporate external textual resources with different aggregation approaches. We conclude the settings of our DNN model in Table 3.

Table 3. Model settings for each run of LADT. The marked runs adopt post-processing.

Run ID	Image	*Visual concept sources*				Concept aggregation	Self-feedback
		Attrib.	Place	Object	Label		
Run01	✓			✓	✓	averaging	
Run02	✓			✓	✓	concept-query rel.	
Run03	✓	✓	✓	✓	✓	concept-query rel.	
Run04	✓		✓	✓	✓	averaging	
Run05	✓					-	
Run06	✓		✓	✓	✓	averaging	✓
Run07	✓		✓	✓	✓	concept-concept rel.	✓
Run08	✓		✓	✓	✓	concept-query rel.	✓
Run09*	✓		✓	✓	✓	concept-concept rel.	✓
Run10*	✓		✓	✓	✓	averaging	

For DNN optimization, we apply Adam optimizer on the binary cross-entropy loss with a learning rate of 1×10^{-5}. The size of mini-batch is 32, and the number of training epochs is at most 40 with early-stopping mechanism. The official result is reported in Table 4, including precision, recall, and F1 score averaged over sixteen activities (official result reports recall score). We observed that the recall of the model increases when adopting adequate concept sets and aggregation strategies, while the precision does not necessarily increase.

Table 4. Official results of LADT. Highest score in each column is shown in bold.

Run ID	Precision	Recall	Micro-F1
Run01	0.7522	0.3840	0.5084
Run02	0.7318	0.4149	0.5296
Run03	0.7261	0.4023	0.5177
Run04	**0.7540**	0.4025	0.5248
Run05	0.7084	0.3606	0.4780
Run06	0.7347	0.4197	0.5343
Run07	0.7535	0.4168	0.5367
Run08	0.7307	**0.4332**	**0.5439**
Run09	0.7532	0.4125	0.5331
Run10	0.7372	0.4116	0.5283

6 Conclusion

The paper presents our approaches to moment retrieval and automated activity annotation for lifelogging. For moment retrieval, we introduce external textual knowledge to reduce the semantic gap between textual queries and the visual concepts extracted by CV models. For automated activity annotation, we incorporate textual features aggregated in an unordered fashion to enrich the training data for supervised DNN models. Experimental results show our strategies achieve better performance compared to naïve baseline models.

Acknowledgments

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-107-2634-F-002-011-, and MOST-108-2634-F-002-008-.

References

1. Gurrin, C., Joho, H., Hopfgartner, F., Dang-Nguyen, D.T., Zhou, L., Healy, G., Albatal, R.: Overview of NTCIR-14 lifelog-3 task. In: Proceedings of NTCIR-14, Tokyo (2019).
2. Tang, T.H., Fu, M.H., Huang, H.H., Chen, K.T., Chen, H.H.: Visual concept selection with textual knowledge for understanding activities of daily living and life moment retrieval. In: CLEF2018 Working Notes (CEUR Workshop Proceedings) (2018).
3. Lin, J., Molino, A., Xu, Q., Fang, F., Subbaraju, V., Lim, J.H.: VCI2R at the NTCIR-13 Lifelog-2 lifelog semantic access task. In: Proceedings of NTCIR-13. Tokyo (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a largescale hierarchical image database. In: proceedings of CVPR (2009).
5. Bolei, Z., Agata, L., Aditya, K., Aude, O., Antonio, T.: Places: a 10 million image database for scene recognition. In: Proceedings of TPAMI (2017).
6. Yamamoto, S., Nishimura, T., Akagi, Y., Takimoto, Y., Inoue, T., and Toda, H.: PBG at the NTCIR-13 Lifelog-2 LAT, LAST, and LEST tasks. In: Proceedings of NTCIR-13, Tokyo (2017).
7. Dogariu, M., Ionescu, B.: Multimedia Lab @ ImageCLEF 2018 lifelog moment retrieval task. In: CLEF2018 Working Notes (2018).
8. Abdallah, F.B., Feki, G., Ezzarka, M., Ammar, A.B., Amar, C.B.: Regim Lab Team at ImageCLEFlifelog LMRT Task 2018. In: CLEF2018 Working Notes (2018).
9. Tran, M.T., Truong, T.D., Dinh-Duy, T., Vo-Ho, V.K., Luong, Q.A., Nguyen, V.T.: Lifelog moment retrieval with visual concept fusion and text-based query expansion. In: CLEF2018 Working Notes (2018).
10. Dogariu, M., Ionescu, B.: A textual filtering of HOG-based hierarchical clustering of lifelog data. In: CLEF2017 Working Notes (2017).
11. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. In: Foundations and Trends in Information Retrieval archive, Vol 3 Issue 4. ACM (2009).
12. Simonyan, K., Zisserman, Z.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of ICLR (2015).
13. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of EMNLP (2014).
14. Iyyer, M., Manjunatha, V., BoydGraber, J., Daum´e III, H.: Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of ACL (2015).
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In proceedings of NIPS (2013).