VTIR at the NTCIR-12 2016 Lifelog Semantic Access Task

Long Xia
Virginia Tech
longxia1@vt.edu

Yufeng Ma
Virginia Tech
yufengma@vt.edu

Weiguo Fan
Virginia Tech
wfan@vt.edu

ABSTRACT

The VTIR team participated in the Lifelog Semantic Access Task of the NTCIR-12 2016. We proposed an approach to pre-process the visual concepts of the photos and reconstruct queries from two aspects. Due to limited time, our efforts were constrained. Our main contribution was to learn and incorporate location of each photo to improve search accuracy. The evaluation demonstrated that our approach achieved pretty decent results with querying simple scenarios, while a more sophisticated model is needed for search complicated scenario, including range search. This report describes our approach to improve retrieval accuracy and discusses the results obtained, and proposes some possible improvements that can be made in the future.

Team Name

VTIR

Subtask

Lifelog Semantic Access Task (LSAT) to explore search and retrieval from lifelogs

1. INTRODUCTION

We participated in Lifelog Semantic Access Task to explore search and retrieval from lifelogs. With the widespread popularity of wearable sensors, people can digitally record their life experience in great details [1], which would generate a rich archive of life experience. Being able to effectively retrieve certain events, or activity from such massive dataset will enable us to develop more useful tools and applications for information access. This task is to retrieve specific events or activity that happened in a lifelogger's life. For example, "Find the moment(s) where I use my coffee machine."

2. RELATED WORKS AND EXPERIMENTS

NDCG (Normalized Discounted Cumulative Gain) at different depths will be applied to evaluate the results. In order to achieve information retrieval purposes, two questions needs to be answered [2]: (1) in such a query-document relevance ranked list, what features of the documents and queries can be used to do the relevance calculation and (2) how to develop a custom ranking relevance scoring function that can effective retrieve photos.

2.1 Feature Expansion

WordNet [3] is widely used to find the sets of cognitive synonyms. It was used in our experiment to expand the visual concepts features.

We first did an analysis of the queries and found out the location is an important component in the information retrieval process. Thus, in the offline phase, we first randomly selected 3000 photos from the dataset and annotated them with pre-defined location labels (see Fig. 1). Recently, encouraged by the significant improvement of deep learning in photo classification [4], we train a neural networks on the training dataset (the first 2000 photos) and then apply the classifier for the testing dataset (the last 1000 photos). The system gave us 80.4% accuracy in determining the locations. Although the accuracy is not great, it is reasonable to incorporate this feature into the each photo. The distribution of the location attributes in our annotated dataset is plotted in Fig. 2.

Recent years, with the development of deep learning, a number of computer vision problems have been improved dramatically, including

Location	Office (1)	Home (2)	Street (3)	Transportation (4)	Meeting Room (5)	Restaurant (6)	Stores (7)	Others (8)
No. of photos	857	405	182	458	365	182	139	412

Fig. 1. Pre-defined location labels and total number of each label in the 3000 randomly selected lifelog dataset

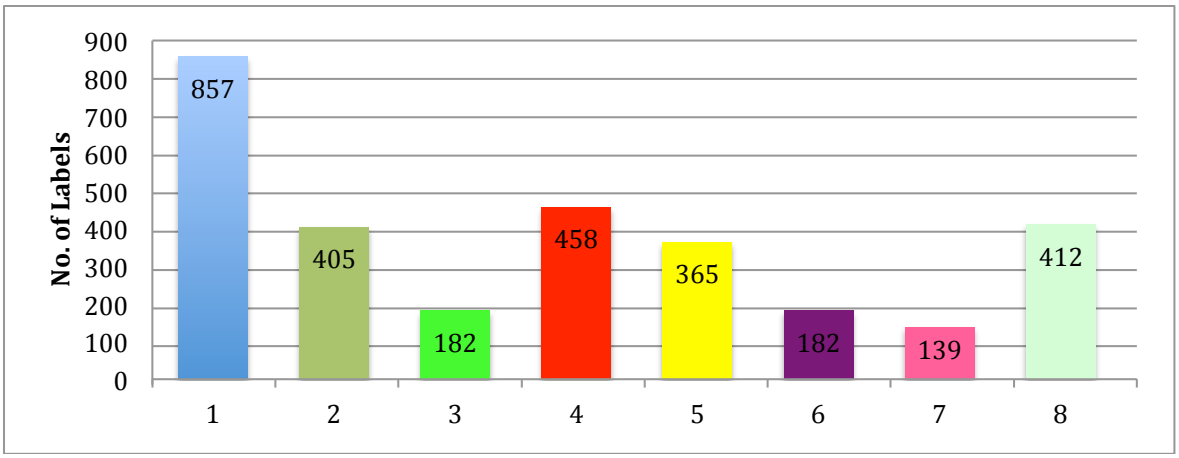

Fig. 2. Distribution for each location attribute in the 3000 randomly selected annotated dataset

image classification, object detection [4], visual recognition [5] and representation [6], etc. These techniques should be able to obtain better features of the photos, including better and more accurate visual concepts, precise description of the content of photos which can be easier retrieved, etc. However, due to limited time we had, we were not able to implement these techniques.

2.2 Search Query Expansion

To improve both precision and recall of our results we make use of query expansion technique. Both automatic and manual methods were applied to achieve the expanding purpose.

For all the 48 formal run queries, we removed all the stop words, followed by manually constructing the queries. Specifically, one human expert analyzed each query and only included meaningful words in that query. For example, "Find the moments when I was repairing my car by replacing the wheel" was converted to "repairing car wheel". Then, WordNet was applied to resulting queries to expand the query. Lastly, the locations of the events were labeled by human manually. For example, for query "Find the moments in which I was in a meeting at work with 2 or more people", the location "Meeting Room" was added.

2.3 Custom Ranking Function

There are two fields in our designed model: one is the visual concept of each photo, and the other one is the location of each event. The proper weights for incorporating these two scores need to be found. These weights are significant because they contribute to the total relevance scores. W_{loc} represents the weightage given to the relevance score from the location matching in the total relevance score. $Score_{loc}$ represents whether the location of the query matches with location of the photo or not. It is a binary value. Number 1 means matched, while 0 means not matched. $Score_{visual}$ represents the relevance score of how the visual concepts of certain photo are relevant to certain query. W_{visual} represents the weightage of $score_{visual}$ in the total relevance score.

Hence our custom ranking relevance scoring function looks like:

Custom Relevance Score = $score_{loc}$ * W_{loc} + $score_{visual}$ * W_{visual}

In order to determine the weights values, we should first develop a model to calculate the $score_{loc}$ and $score_{visual}$. $Score_{loc}$ is a binary number which is straightforward. To develop the best model for calculating $score_{visual}$, we tested different models, and Okapi BM25 turned out to be the best one based on our manual evaluations.

To determine the best values of these weights, W_{loc} was fixed to be 1 and different W_{visual} values [0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30] were tested. We use a set of queries in the queries set that represent a good coverage. We run these queries against our system and collect the search results. The top 15 results were evaluated manually and the system achieved best results when W_{visual} = 0.3. Thus, the custom ranking function applied in our retrieval system is:

Custom Relevance Score = $score_{loc}$ * 1 + $score_{visual}$ * 0.3

3. EVALUATIONS AND CONCLUSIONS

3.1 Evaluations

This year's LSAT task has 48 evaluation topics (001-048). This section summarizes our submitted run and analyzes results.

3.1.1 Submitted Run

Due to time limit, we only submitted one search run (Yufeng-run_1.csv). Fig. 3 summarized our results.

3.1.2 Results Analysis

According Fig. 3, our model did not achieve very promising overall results. However, after carefully analyzing the specific results of each queries, we figured out that our system achieved good results on some queries searching for simple scenarios. For examples, for query 24, "Find the moment(s) when I was repairing my car by replacing the wheel.", we achieve NDCG score of 1, and for query 20, "Find the moment(s) when I was taking a photo of an Airbus A380 airplane.", we achieved NDCG score of 0.63. This indicates our model did work well on these scenarios. However, for queries searching for complicated scenarios, such as range search, we often got NDCG score of 0, indicating our model did not work well on these cases. Another interesting finding is that for some queries search, the top ranked photos returned were about completely different subjects and obejcts, one possible reason for this may be that the visual concepts provided might not be accurate.

3.2 Conclusions and Perspective

In this paper, we demonstrated the lifelog semantic access problem from information retrieval perspective and described our approach for NTCIR-12 2016 Lifelog. We presented two methods for features and queries expansion, and applied Okapi BM25 in calculating the relevance score. We were running out of time to design and implement a more sophisticated for this year's competition, and we did not got a decent overall accuracy. However, we proposed some strategies to make improvements for next year's competition. Firstly, we plan to do some annotations to each picture. By adding this feature, we expect the performance of range search (e.g. Find the moment(s) when I boarded a Red taxi, only to get out again shortly afterwards). Also, we plan to do the topic modeling on each picture and also add those features to the picture. Furthermore, with the development of convolutional neural network in computer vision, we plan to incorporate it in our future model to do visual recognition and representation on all the photos to further expand the features. We expect these should improve the general retrieval performance to some degree.

	Run	NDCG	MAP	P@10	R-prec
Event Level	Yufeng-run_1.csv	0.073	0.050	0.027	0.040
Image Level	Yufeng-run_1.csv	0.024	0.010	0.054	0.014

Fig. 3. Mean of evaluation results over all the queries for NDCG, MAP, P@10 and R-prec

4. REFERENCES

[1] Gurrin, Cathal, Alan F. Smeaton, and Aiden R. Doherty. Lifelogging: Personal big data. *Foundations and trends in information retrieval*, pages 1-125, 2014.
[2] Xu, Tan, McNamee, Paul, Orad, Douglas W. HLTCOE at TREC 2014: Microblog and Clinical Decision Support. *TREC*. 2014.
[3] Christiane Fellbaum. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. 1998.
[4] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv*, 2015.
[5] Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625-2634. 2015.
[6] Chen, Xinlei, and Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv*, 2014.