Visual Concept Selection with Textual Knowledge for Understanding Activities of Daily Living and Life Moment Retrieval

Tsun-Hsien Tang[12*], Min-Huan Fu[1*], Hen-Hsen Huang[1], Kuan-Ta Chen[2] and Hsin-Hsi Chen[13]

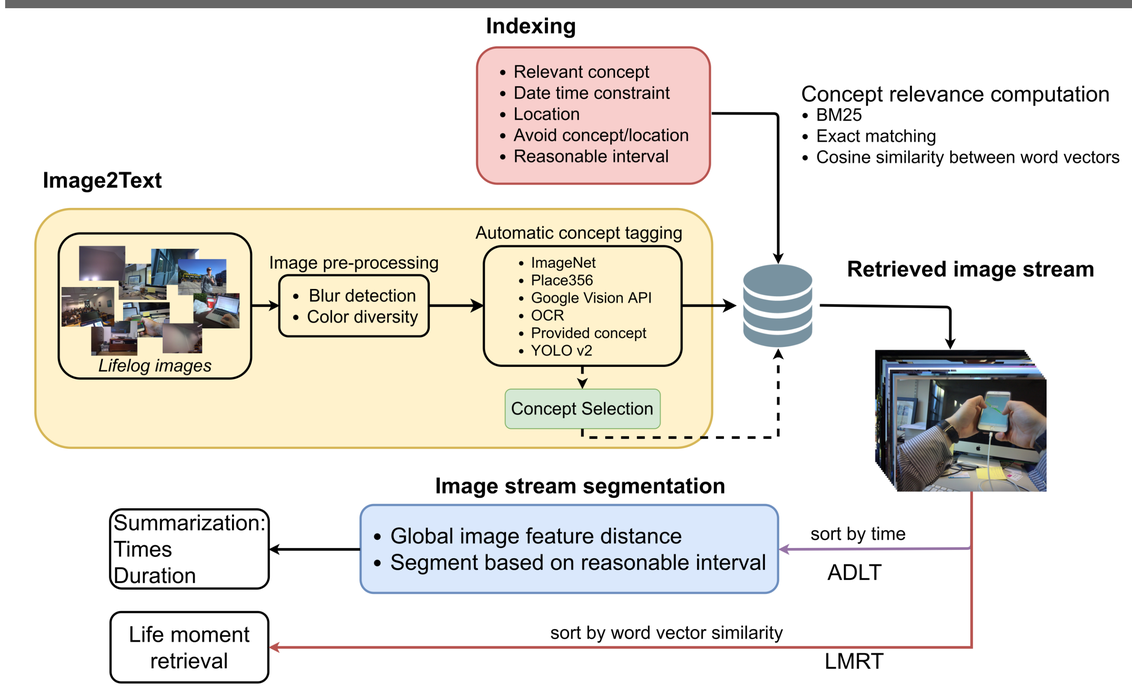
[1] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[3] MOST Joint Research Center for AI Technology and AII Vista Healthcare Taiwan
{ thtang, mhfu, hhhuang}@nlg.csie.ntu.edu.tw; swc@iis.sinica.edu.tw; hhchen@ntu.edu.tw

Motivation

- Personalized multimodal recording, together with dedicated lifelogging applications for smartphones, becomes more popular nowadays.
- However, numerous personalized data that are acquired, recorded, and stored still remain challenging to access by their owners.
- Therefore, a system for recapping precious life moments is needed.



Retrieval Framework

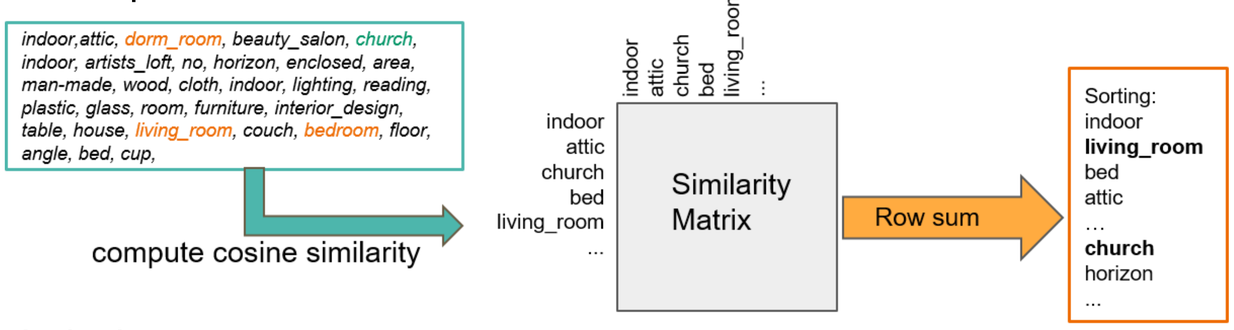


Method

1. Visual Concept Extraction:
   - Prune low quality images with blurriness and color diversity detection
   - Automatically extract concept from image using pre-trained network.

Provided concept: ['indoor', 'man', 'person']
ImageNet concept: ['patio', 'torch', 'bucket', 'Christmas_stocking', 'lab_coat']
Place365 concept: ['indoor', 'airplane_cabin', 'fastfood_restaurant', 'socializing', 'working']
Object detection: ['person', 'cup', 'person']
Google vision api:
label: ['car', 'man', 'fun', 'male', 'vehicle', 'recreation', 'vacation', 'tree']
text detection: ['cafe', 'the']

   - Concept Selection

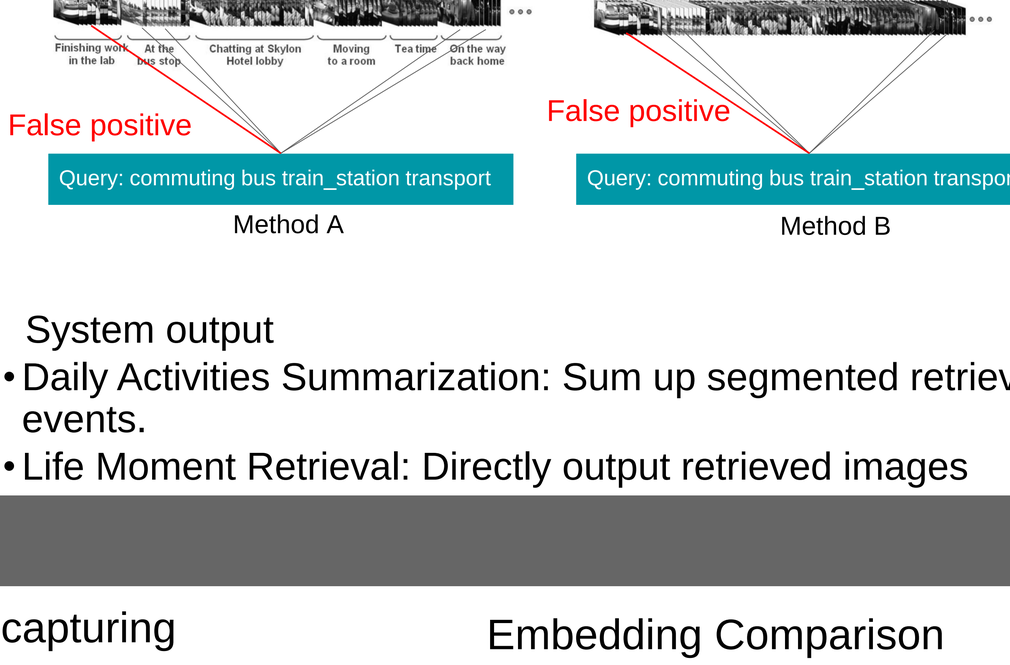


2. Indexing
   - Metadata Preprocessing
   - Retrieval model in NLP: Exact Matching, BM25, Word Embedding

3. Image Stream Segmentation
   - Deep Learnt Feature: Compute distance of two vectors.
   - Event Interval: Set a reasonable duration for a specific activity.

The L2 distance of **(1)Autoencoder latent space (2)FC7 features** between each image.

1. manual set ***reasonable interval***
2. count adjacent "minutes" as *interval*
3. if *interval* < *reasonable interval*:
   - merge images as event



4. System output
   - Daily Activities Summarization: Sum up segmented retrieved events.
   - Life Moment Retrieval: Directly output retrieved images

Experiment Results

1. Performances in ADLT

Percentage Dissimilarity:

$$ADL_{score} = \frac{1}{2}\left(max\left(0, 1 - \left|\frac{n - n_{gt}}{n_{gt}}\right|\right) + max\left(0, 1 - \left|\frac{m - m_{gt}}{m_{gt}}\right|\right)\right)$$

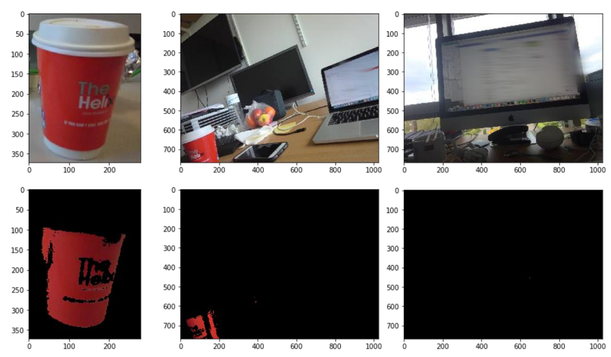
n : how many times
m : how long (in minutes)
n_{gt} and m_{gt} denote ground truth

Pipeline	Percentage Dissimilarity
vanilla retrieval	0.2434
fine-tuned query	0.2850
fine-tuned query + concept selection (N)	0.3850
fine-tuned query + concept selection (G) + coffee capturing	0.4592
fine-tuned query + concept selection (N) + coffee capturing	**0.4787** (Rank 2)

2. Performances in LMRT

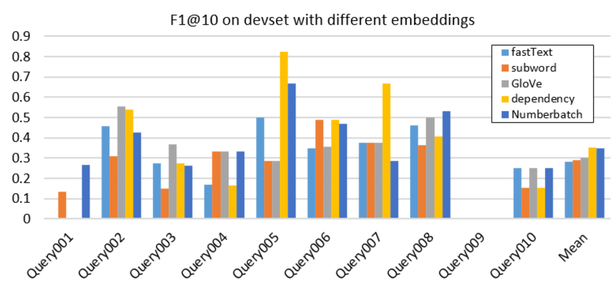
Run ID	Description	F1@10
Run 1	automatic query (baseline)	0.177
Run 3	automatic query + constraints	0.223
Run 4	fine-tuned query + fine-tuned constraints	**0.395** (Rank 4)
Run 5	fine-tuned query + fine-tuned constraints + clustering	0.354

Coffee capturing



We also demonstrate how image feature (RGB here) could help user to retrieve image more precise.

Embedding Comparison



The word embeddings that associate with additional contextual information such as syntactic dependency or lexical ontology provides better performance.

Conclusions

In both subtasks, we introduce the external textual knowledge to reduce the semantic gap between the user query and the visual concepts extracted by the latest CV tools. Experimental results show that filtering out noisy concepts could significantly improve the performance. Besides, proper human intervention for query refinement makes the retrieval more precise.

* Equal Contribution

Reference:
https://www.engadget.com/2018/02/27/google-clips-ai-camera-review/
https://www.digitaltrends.com/digital-camera-reviews/gopro-hero-2018-review/
https://www.youtube.com/watch?v=D3BJ5bt5VbA