PKU_ICST at TRECVID 2017: Instance Search Task

Yuxin Peng, Xin Huang, Jinwei Qi, Junchao Zhang, Junjie Zhao, Mingkuan Yuan, Yunkan Zhuo, Jingze Chi, and Yuxin Yuan
pengyuxin@pku.edu.cn

PEKING UNIVERSITY

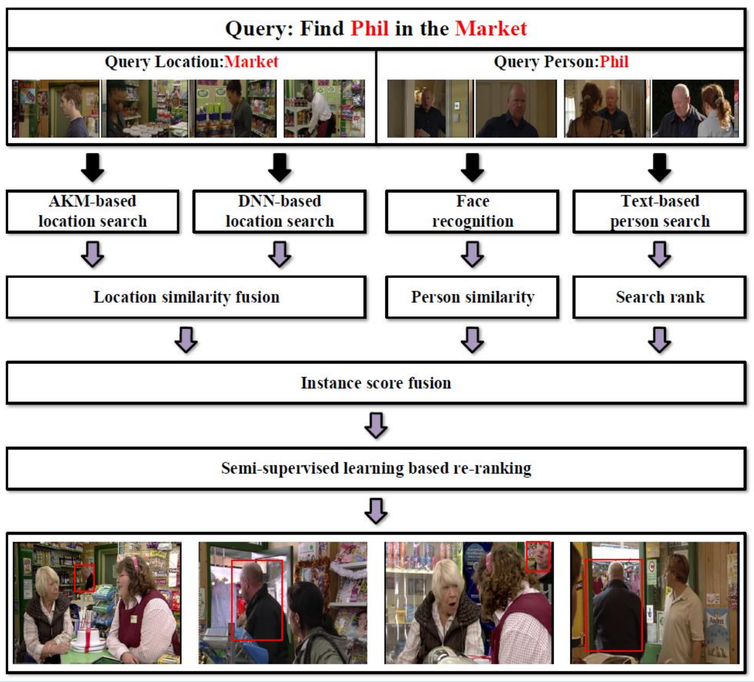
多媒体信息处理研究室
Multimedia Information Processing Lab (MIPL)
Institute of Computer Science&Technology, Peking University

Overview

We ranked ***1st*** in both ***automatic*** and ***interactive search*** of Instance Search (INS) tasks.

The approach consists of two stages:

- **Similarity computing**: *Location-specific search* and *person-specific search* are conducted and fused
- **Result re-ranking**: *Semi-supervised re-ranking method* is applied to filter noisy shots in top-ranked results



Location-specific Search

We conduct location-specific search with both ***handcrafted*** and ***deep*** features:

- Similarity score of *AKM-based search*:

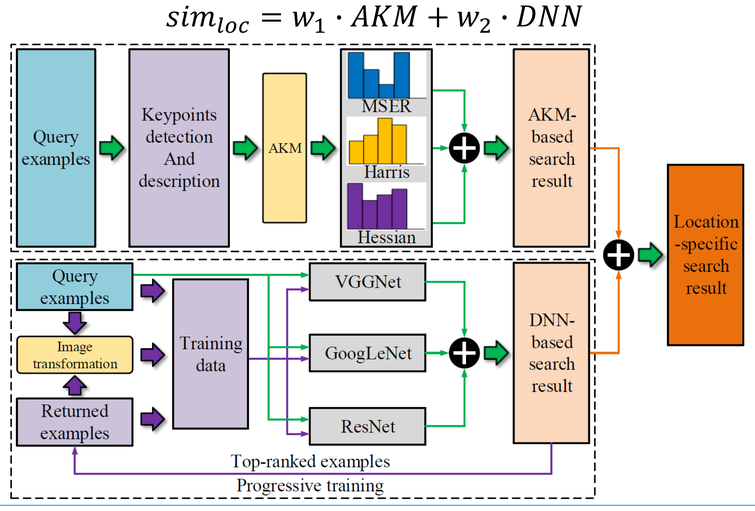
$$AKM = \frac{1}{N}\sum_{k} BOW^{(k)}$$

- Similarity score of *DNN-based search*:

$$DNN = \frac{1}{3}(VGG + GOOGLE + RESNET)$$

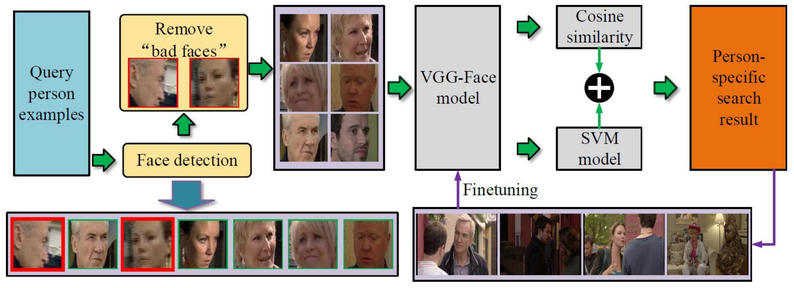
- Combination:

$$sim_{loc} = w_1 \cdot AKM + w_2 \cdot DNN$$



Person-specific Search

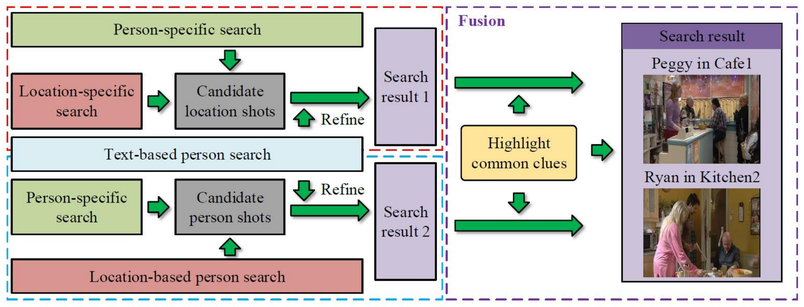
- We conduct person-specific search by ***deep face recognition*** method.
  - Remove "**bad faces**" to filter noise of query
  - Integrate cosine similarity and SVM prediction scores to get the search result.



- We also conduct text-based person search.
  - Persons' relevant information from publicly available sources was mined to get the search result.

Fusion and Re-ranking

Fusion: Combine the ***location*** and ***person*** clues from two different directions.



Re-ranking: Filter noise based on semi-supervised learning.

- Obtain affinity matrix W of top-ranked shots F:

$$W_{ij} = \begin{cases} \frac{F_i^T \cdot F_j}{|F_i| \cdot |F_j|}, i \neq j \\ 0, \quad i = j \end{cases}, \ i,j = \{1,2,\cdots,n\}$$

- Update W according to k-NN graph:

$$W_{ij} = \begin{cases} W_{ij}, F_i \in KNN(F_j) \\ 0, \ otherwise \end{cases}, \ i,j = \{1,2,\cdots,n\}$$

- Construct the matrix:

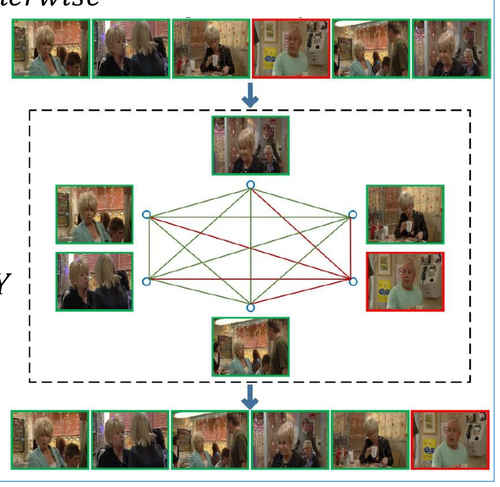
$$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

where D was a diagonal matrix

- Re-rank search result:

$$G_{t+1} = \alpha S G_t + (1-\alpha) Y$$

where Y was the ranked list obtained by above **fusion** step



Results and Conclusions

Type	ID	MAP	Brief description
Automatic	RUN1_A	0.448	AKM+DNN+Face
	RUN1_E	0.471	AKM+DNN+Face
	RUN2_A	0.531	RUN1+Text
	RUN2_E	0.549	RUN1+Text
	RUN3_A	0.528	RUN2+Re-rank
	RUN3_E	**0.549**	RUN2+Re-rank
Interactive	RUN4	**0.677**	RUN2+Human feedback

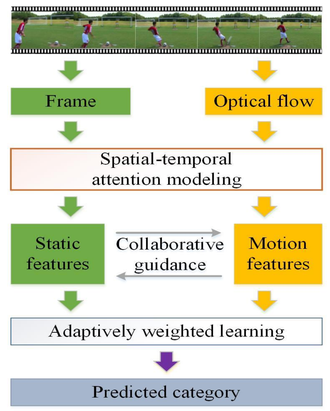
- **Video examples** are helpful for accuracy improvement
- Automatic removal of "**bad faces**" is important
- Fusion of **location and person similarity** is a key factor of the instance search

Our Related Works

Welcome to our website for papers and source codes: *http://www.icst.pku.edu.cn/mipl/*

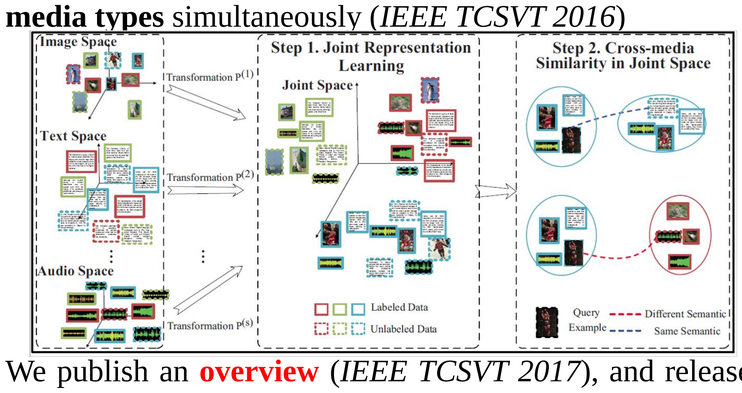
Video concept recognition:

- We propose spatial-temporal collaborative learning method (*arXiv:1711.03273*):
  - Jointly model spatial and temporal attention
  - Mine complementary clues of static and motion information



Cross-media Retrieval:

- Retrieve across **different media types**, such as image, text, audio and video
- We propose the **first work** of cross-media retrieval for **5 media types** simultaneously (*IEEE TCSVT 2016*)



- We publish an **overview** (*IEEE TCSVT 2017*), and release a large-scale dataset ***PKU-XMediaNet*** with 5 media types

Fine-grained image analysis:

- To recognize **hundreds of subcategories** under basic-level categories, like "dog" and "bird"
- We propose the **first work** without **object or parts annotations** in training and testing phases (*IEEE TIP 2017*)