北京邮电大学
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

BUPT-MCPRL at Trecvid2014 Instance Search Task

Wenhui Jiang (jiang1st@bupt.edu.cn)

Zhicheng Zhao, Qi Chen, Jinlong Zhao, Yuhui Huang,

Xiang Zhao, Lanbo Li, Yanyun Zhao, Fei Su, Anni Cai

MCPR Lab

Beijing University of Posts and Telecommunications

Our submission

- BOW baseline + CNN as global feature: 22.7%

  CNN as global feature boosts the performance by about 3% (estimated in INS2013).

- BOW baseline + Query expansion + CNN as global feature: 22.1 %

  That's not normal. We are investigating on it.

- BOW baseline + Localized CNN search : 21.6%

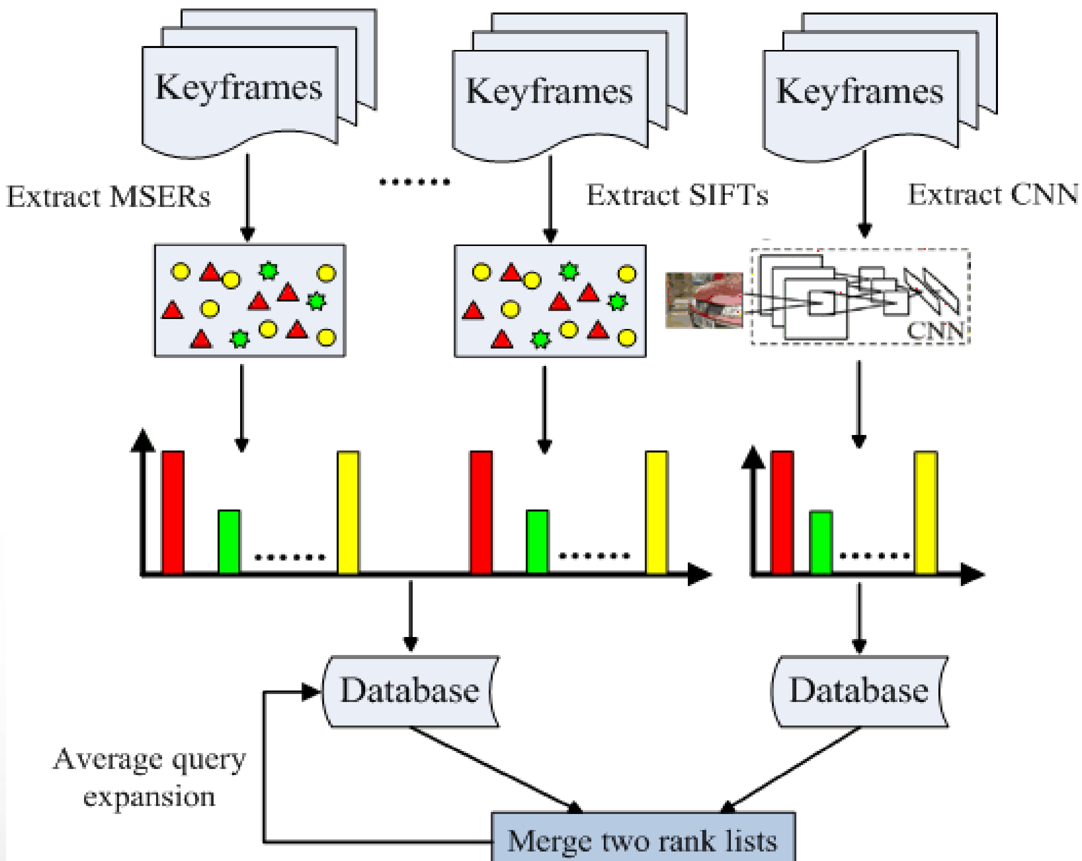
  Localized CNN search boosts the performance by about 0.5%.

- Interactive Run: BOW baseline + Query expansion (Interactive): 23.8%

Brief introduction

- **Reference Dataset**
  - 470K shots
  - 2 key frames per second
  - Max pooling for shot score
- **Query Images**
  - Average pooling for query score
- **Feature Model**
  - Bag-of-words
  - Convolutional neural networks

System Overview



BOW Highlights

- Three kinds of local features + BOW framework

  + ≈9% mAP
- Contextual weighting

  + ≈3% mAP
- Burstiness

  + ≈2% mAP

Three kinds of local features

- Hessian detector + RootSIFT (128D)
- MSER detector + RootSIFT (128D)
- Harris Laplace + HsvSIFT (384D)
- AKM for training codebook of size 1M

local features	points per image	mAP(INS2013)
MSER + RootSIFT	around 150	16.308
Hessian + RootSIFT	around 500	12.739
Harris + HsvSIFT	around 250	12.967
Total	around 900	21.731

Rich features are important, because they are complementary.

Contextual weighting

- Set different weights on ROI and backgrounds: In the aspect of metric

Typical scheme: $$sim(q,d) = \sum_{i=1}^{D} \alpha_i q_i d_i \quad \text{, where } \alpha_i = \begin{cases} \beta \ (\in ROI) \\ 1 \ (\notin ROI) \end{cases} \qquad (1)$$

Query: {2,1,1} Image 1: {2,0,3} Image 2: {2,1,1}

Similarity (take inner product and L2-normalization as an example, and set β=3):

$$sim(Q, I_1) = 1.47$$
$$sim(Q, I_2) = 1.33$$

Contextual weighting

- A good similarity measurement include of consistent:
  - Similarity kernel.
  - Normalization scheme.
- Good similarity measurement satisfies:
  - Self-similarity equals to one;
  - Self-similarity is the largest.
- L2-norm + inner product √

  L1-norm + inner product ×
- Advise:
  - When you want to set larger weights on ROI descriptors, you may also need to modify the normalization scheme.

Boost the mAP by 3%

Burstiness

Definition: A visual word is more likely to appear in an image if it already appeared once in that image. [Jegou. CVPR 2009]

- If we first normalize the feature vector, then calculate the similarity : image with very few descriptors equals to the image contains several dominant descriptors. This also leads to burstiness.
- Advise: L1-based similarity kernel rather than L2-based.

Boost the mAP by 2%

What’s next?

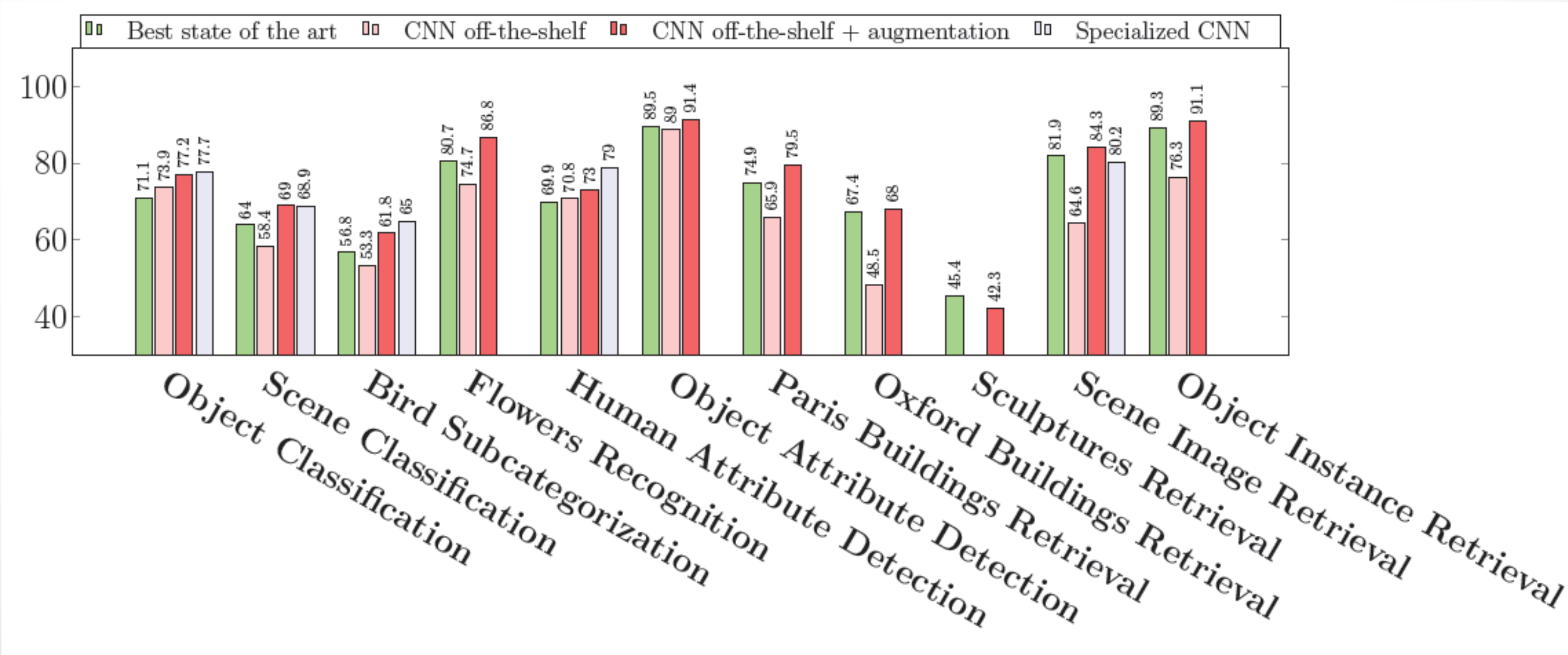
- Local features are **unable** to solve
  - — Smooth objects or objects are more suitable to describe using shape etc.
  - — Small objects which could extract few local features

- What’s next?
  - — Introduce better similarity measurement?
  - — Keep ensembling more features?

北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

What's next?

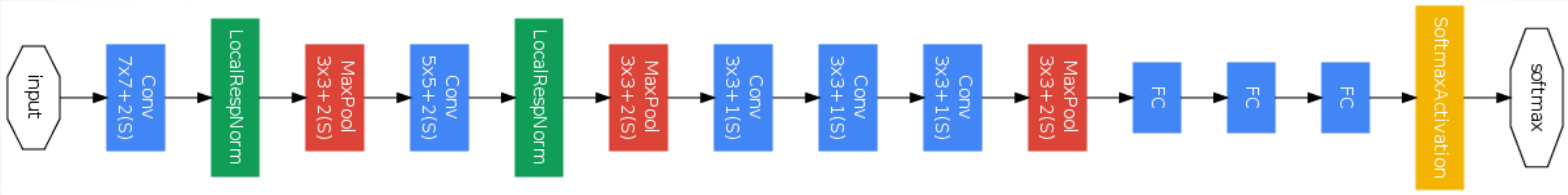
- How well would Deep Learning work for instance search?



[Razavian et al. CVPRw 2014]

Convolutional neural network

- Decaf has shown that CNN trained on ImageNet2012 1000CLS has good generalization.



[Krizhevsky et al. NIPS 2012]

Convolutional neural network

- **Two schemes**
  - As global features
    - + ≈3% mAP
  - Generic object detection + CNN
    - + ≈1% mAP

Convolutional neural network

- **Scheme 1: As global features**
  - Activations from a certain layer as global features.
  - CNN takes the entire image as the input, therefore it is unable to emphasize the ROI.
  - Relatively strict geometric information

Layer	Dim	Metric	mAP (using CNN only)
Fc6	4096	L2	3.84
Fc6 + Relu	4096	SSR	3.43
Fc7 + Relu	4096	L2	3.07
Fc7 + Relu	4096	SSR	2.67
Fc8	1000	SSR	1.34

Boost the mAP by 3% (combined with BOW)

Convolutional neural network

- **Scheme 2: Localized search**
  - — Instance search is inherently asymmetric.
  - — CNN is not like BOW, it has fewer geometric correspondences, especially for the output of fully connected layer.

- How to deal with the asymmetric problem of CNN?
  - — Train a specific CNN

    But where is the training set come from?

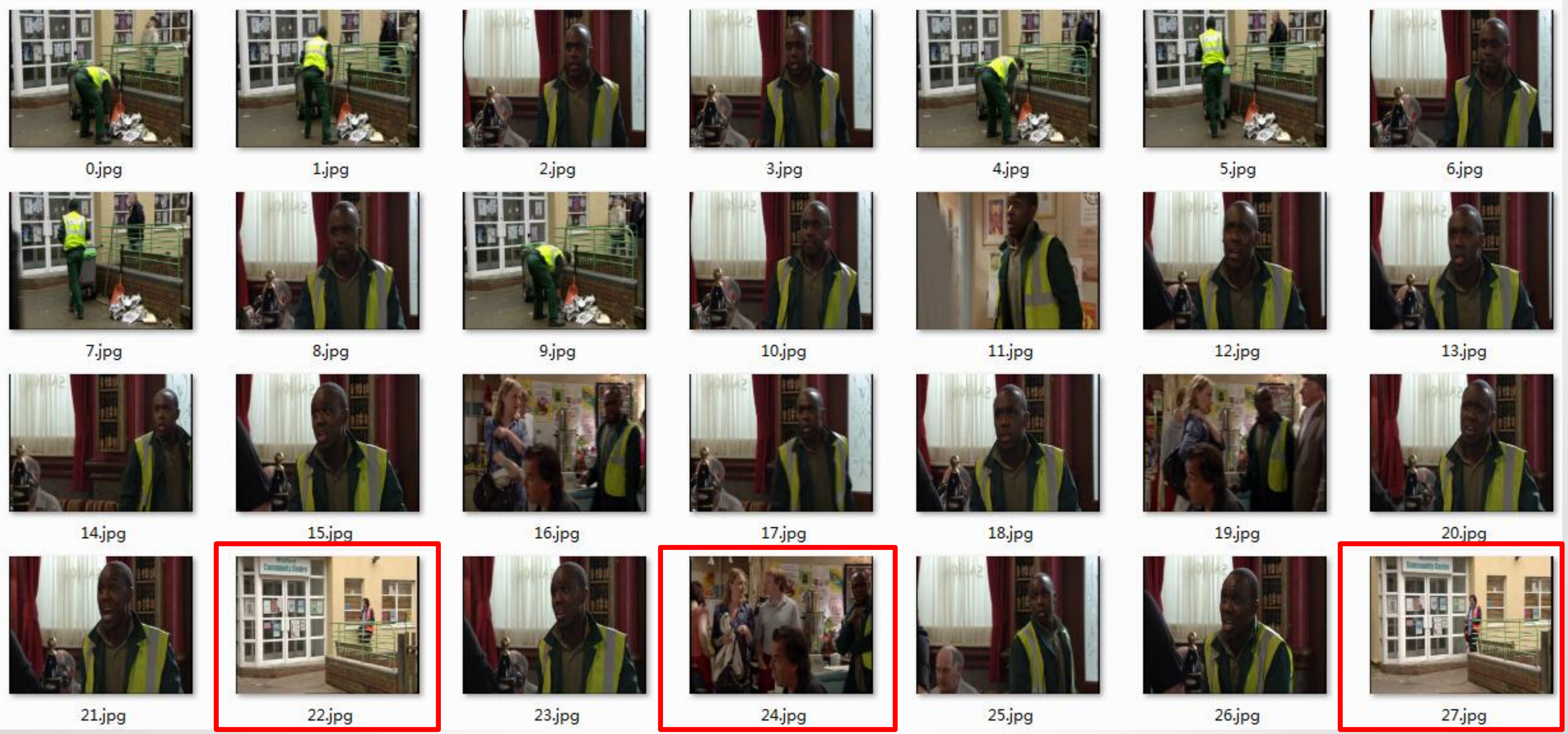
  - — Generic object detection (derived from RCNN) + CNN feature comparison

    Problem: Designing an efficient indexing system is important.

    As a trial run, we only use it for reranking the top 100 results.

Boost the mAP by 1%

Topic 9113, result from BOW baseline. Images in red box are false results。



Topic 9113, result after reranking。


0.jpg


1.jpg


2.jpg


3.jpg


4.jpg


5.jpg


6.jpg


7.jpg


8.jpg


9.jpg


10.jpg


11.jpg


12.jpg


13.jpg


14.jpg


15.jpg


16.jpg


17.jpg


18.jpg


19.jpg


20.jpg


21.jpg


22.jpg


23.jpg


24.jpg


25.jpg


26.jpg


27.jpg

Failure examples

1.jpg

2.jpg


3.jpg


4.jpg


5.jpg

6.jpg


7.jpg


8.jpg


9.jpg


10.jpg


11.jpg


12.jpg

Failure examples: After reranking

Problems

- **The input region is limited to a rectangle, not arbitrary shape.**

Problems

Instance Search	Object Detection
1. No suitable training data;	1. Enough training data;
2. Focus on **both intra-class and inter-class** analysis;	2. Mainly focus on **inter-class** analysis;
3. Objects to be retrieved could be anything;	3. Object class to be detected is specified ahead of time;
4. Require real-time response.	4. Could be performed off-line.
5. Focus on finding relevant image from a large dataset.	5. Focus on detecting relevant object in a given image.

Thanks!

jiang1st@bupt.edu.cn
https://sites.google.com/site/whjiangpage/

http://www.bupt-mcprl.net