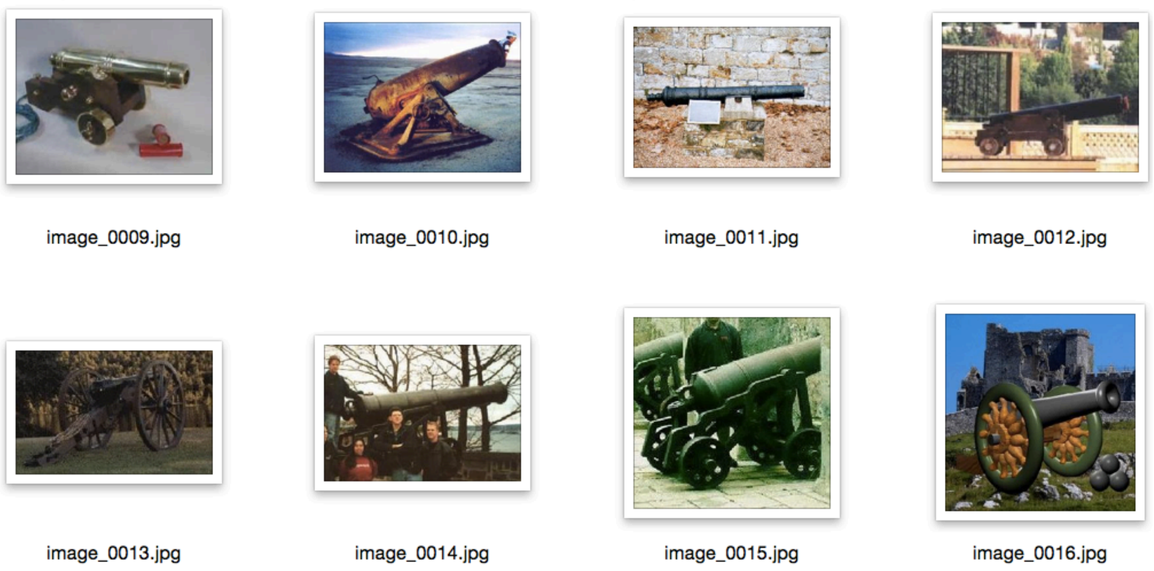
SS4-40: A Fast Method of Visual Words Assignment of Bag-of-Features for Object Recognition

Meng-Jiun Chiou[†1,†2] Toshihiko Yamasaki[†1] Kiyoharu Aizawa[†1]
[†1]The University of Tokyo [†2]National Chiao Tung University

Background

Recently, with rapid growth of smart devices, it's easy for people to upload and share their photos.



∗ Problem

However with a **large** amount of query pictures, how to **classify them into right category fast**?

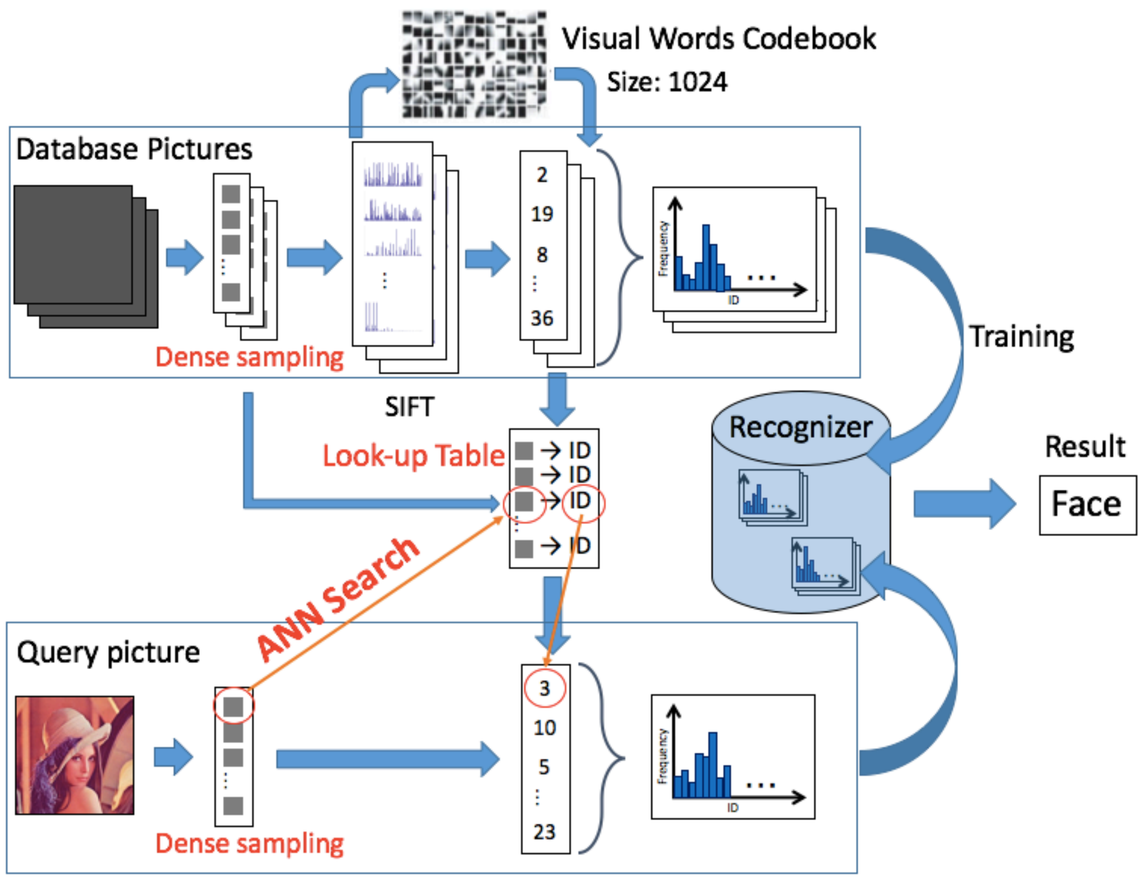
∗ Investigation: The time needed for different operations of mobile image retrieval [1]

Client	
Operation	**Time**
Feature Extraction	1000-1500 ms
Search Visual Words	37 ms
Encode Visual Words	4 ms

one the most time-consuming parts

∗ Goal: Trying to improve the speed of large-scale object recognition problem **without feature extraction**.

Proposed Method



Proposed Method = SIFT (Database only) + Dense Sampling + Bag-of-Features + **Look-Up Table** + Classification

16 × 16 pixel patches of database pictures

Corresponding visual words assigned from the visual words codebook

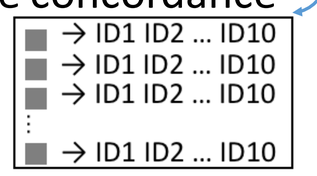
Experiment

∗ Dataset: Caltech 101/256

Property \ Dataset	Caltech 101	Caltech 256
Number of Images	9146	30607
Number of Categories	102 (101+1)	256

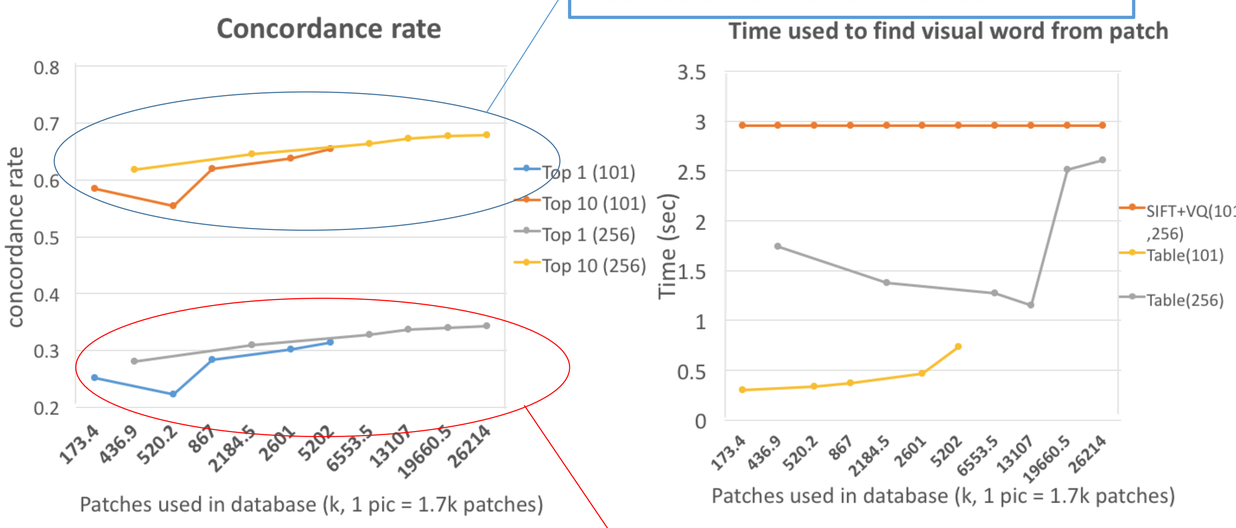
∗ Setting

- Pictures are preprocessed into **gray scale** and **normalized**
- Adopting **FLANN** [2] as the searching algorithm
- Extending the look-up table into [patch -> **10** of the nearest visual words], in order to calculate concordance between them.



∗ Result of Concordance

the nearest visual word found by FLANN is **included in the nearest 10** visual words



the nearest visual word found by FLANN is **the same as the nearest** visual word

∗ Result of Classification (Conventional Method VS Proposed Method)
(Support Vector Machine -- **LIBSVM with RBF-kernel** used)

SVM Classification	Dataset: Caltech101				
M/N	1530/510	1530/1530	3060/510	3060/1530	3060/3060
Accuracy (Conventional Method,%)	**43.64**		**53.48**		
Accuracy (Proposed Method,%)	**29.7**	**27.18**	**30.73**	**37.24**	**31.55**

(M, N is the number of pictures used to train database and form table, respectively.)

Conclusion

With proposed method:

- Time of assigning visual word can be reduced by **up to 85%**
- **Lower** accuracy was observed compared to conventional solution
- Patches can may be revised by **adding some processing in SIFT** (e.g. orientation) to show the distinctness

Reference

[1] Wu Yue, et al. "Local visual words coding for low bit rate mobile visual search." ACMMM, 2012.

[2] Marius Muja and David G. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration", in VISAPP 2009.