Learning a Fine Vocabulary

Andrej Mikulik, Michal Perdoch, Ondrej Chum, Jiri Matas
Center for Machine Perception, Czech Technical University in Prague, Czech Republic

Introduction

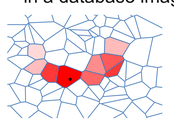
- Context: Large-scale image retrieval based on bag-of-visual-words.
- Euclidean distance on SIFT descriptors reflects patch similarity only locally (small distance implies similarity).
- Fine vector quantization (small cluster cells) separates similar patches into different visual words.
- Idea: Fine-grained clustering followed by learning of groups of similar clusters.
- A novel similarity function (on SIFT descriptors) based on a probabilistic relationships (PR) of descriptors is learned in an unsupervised manner.

Fine Vocabulary with PR Similarity

- Fine (over)partitioning of the SIFT descriptors space.
- 16M words obtained by hierarchical (2 levels, branching factor 4096) approximate k-means.
- The PR similarity is proportional to the probability

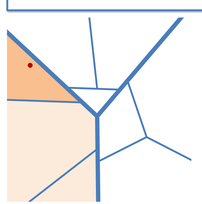
$$P(w_j|w_q)$$

i.e. that an observed word w_q in the query becomes a visual-word w_j in a database image.

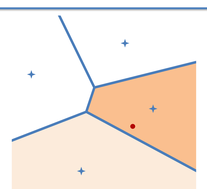


Probabilistic relationship similarity defined in terms of a set of alternative words learned from automatically established correspondences within an image collection.

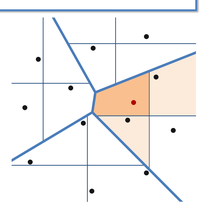
Related Work



Hierarchical scoring [3]
The soft assignment is given by the hierarchical structure.



Soft clustering [5]
Assigns features to r nearest cluster centers.

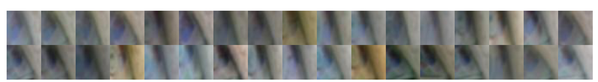


Hamming embedding [2]
Each cell is divided into orthants by a number of hyperplanes. The distance is measured by the number of separating hyperplanes.

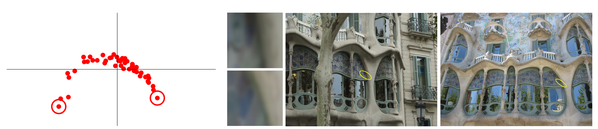
Mining for Sets of Corresponding Patches

- Patch: Hessian-affine point described by SIFT descriptor [4].
- Clusters of spatially related images in the database found using min-hash (the result is a tree structure of each image cluster with affine transformation along edges) [1].
- To avoid wide-baseline matching between every pair in the cluster, 2k-connected circulant graph is constructed on each sub-tree of height two (k set to 10).

Examples of images in a cluster. Yellow ellipses show one set of corresponding – geometrically consistent – patches.



Feature track: a set of corresponding patches from all images in a cluster, treated as a set of projections of a single 3D surface patch.



2D PCA projection of the SIFT descriptors and appearances of the two most distant patches in images. The average SIFT distance within the track is 278, the maximal distance is 591.

- Probability $P(w_j|w_q)$ is estimated from feature tracks as:

$$P(w_j|w_q) \approx \sum_{\mathcal{Z}} \underbrace{P(w_j|z_i)}_{\substack{\text{Probability that a patch from track } z_i \\ \text{is assigned to visual word } w_j}} \underbrace{P(z_i|w_q)}_{\substack{\text{Probability of track (3D patch) } z_i \text{ when} \\ \text{observing visual word } w_q \text{ in an image}}}$$

where z_i is an identifier of a feature track.

- The set of alternative words is defined as:

$$\mathcal{L}=\{w_j|P(w_j|w_q)>0\}$$

- In experiments, we use a subset of $\mathcal{L}$ with at most L top weighted alternative words (L $\leq$ 16).

Training Data

Word probabilistic relationships learnt from:

5,600,000	Flickr images (F5M), images from Oxford 105k and their duplicates explicitly removed
20,000	Clusters of 750,000 images were found using min-hash
1,120,000,000	Features in geometrically verified correspondences
111,000,000	Feature tracks were established
12,300,000	Features appeared in more than 5 images

Image Retrieval Results

method parameters	F5M 16M std	F5M 16M L5	F5M 16M L16	Paris 1M std	Paris 1M SA 3NN
plain	0.554	0.650	**0.674**	0.574	0.652
query expansion	0.695	0.786	**0.795**	0.728	0.772

Mean average precision for selected vocabularies on the Oxford 105k data-set

- **F5M 16M** vs. **Paris 1M std**: 1M dictionary outperforms the 16M without links. For the 0–∞ metric, the 16M visual word dictionary is too fine.
- **F5M 16M std** vs. **16M L5,L16**: Alternative words significantly improve mean average precision.
- **F5M 16M L16** vs. **Paris 1M SA 3NN**: PR similarity function outperforms soft-assignment.

dataset \ method parameters	F5M 16M std	F5M 16M L16	F5M 16M QE	F5M 16M L16 + QE	Philbin et al. ECCV 2010
Oxford 5k	0.618	0.742	0.740	**0.849**	0.707
Oxford 105k	0.554	0.674	0.695	**0.795**	0.615
Paris	0.625	0.749	0.736	**0.824**	0.689
Paris + Oxford 100k	0.533	0.675	0.659	**0.773**	N/A
INRIA holidays -rot	0.742	0.749	0.755	**0.758**	N/A

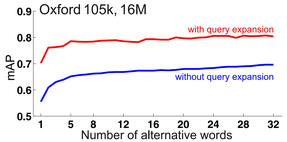
- **F5M 16M L16 + QE** achieved state-of-the-art results.
- **F5M 16M L16** outperforms non-linear projection of SIFT space [Philbin et al. ECCV10].

Query speed: dataset \ method parameters	F5M 16M std	F5M 16M L5	F5M 16M L16	Paris 1M std
Oxford 105k	0.071	0.114	0.195	0.247

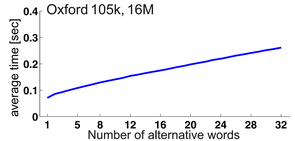
Average execution time per query in seconds (without query expansion)

- **16M** vs. **1M**: The shorter average length of inverted files positively affects average query execution time.

Number of Alternative Words



The quality of the retrieval, expressed as mean average precision (mAP), increases with L, the number of alternative words.



Speed-accuracy trade-off is controllable via L, the number of alternative words.

Conclusions

- The PR similarity increases significantly the accuracy of retrieval, both with and without query expansion.
- The PR similarity outperforms soft assignment in terms of precision.
- The PR similarity (mAP 0.795) outperforms the Hamming embedding approach combined with query expansion, Jegou et al. [2] report the mAP of 0.692 on Oxford 105k.
- The mAP result for 16M L16 is superior to any result published in the literature on the Oxford 5k, the Oxford 105k, and the Paris dataset.
- The similarity measure requires only a constant extra space (independent of the number of images in the database) in comparison with the standard bag-of-words method.
- Retrieval with the proposed similarity function is faster than reference method [4] (using comparable implementations).

References

[1] Chum, O., Matas, J.: Large-scale discovery of spatially related images. PAMI 2010
[2] Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. ECCV 2008
[3] Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. CVPR 2006
[4] Perdoch, M., Chum, O., Matas, J.: Efficient representation of local geometry for large scale object retrieval. CVPR 2009
[5] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. CVPR 2008

The authors are grateful for the support from EC project ICT-215078 DIPLECS, Czech Government under the research program MSM6840770038, GAČR project 102/09/P423, and Google.