Computer Vision 1, Master AI
Tutorial Lecture 4: Bag-of-Words
With Answers

Thomas Mensink and Theo Gevers

1 Local Features

Consider the following image path (I):

$I_{\text{A}} =$

1	1	1	1	1
1	1	1	1	0
1	1	1	0	0
1	1	0	0	0
1	0	0	0	0

1.a Compute the gradient G_x and G_y, using image filters. Which filters do you use?

Answer: $F_x = [1, 0, -1]$ and $F_y = F_x^\top$. Does it matter if we use cross-correlation or convolution? No, in this case it is only a sign flip. The ideas are the same. Could we also use $F_x = [1, -1]$? Yes, the ideas are the same, but be consistent.

1.b Compute the gradient magnitude

Answer: $M = \sqrt{(G_x)^2 + (G_y)^2}$

1.c Compute the gradient orientation (in degrees)

Answer: $\theta = \arctan \frac{G_y}{G_x} \frac{180}{\pi}$ **Note:** you need to make some assumptions,

$$\frac{0}{0} \equiv 0, \qquad \frac{1}{0} \equiv \inf, \qquad \frac{-1}{0} \equiv -\inf,$$

, which results in gradient orientation of 0, 90, and -90 degrees.

1.d Compute the HoG descriptor, using a 9 bin histogram

Answer: Combine the direction of the gradient with its magnitude. So you can differentiate between a ”0” because of no gradient and a ”0” because of a vertical gradient. For real HoG descriptors some more *tricks* are performed, including *unsigned gradients* (using 0 - 180 only), and sharing gradients over bins (ie 30 degree, will count in bin of 0 and bin of 40), these fall beyond the scope of this exercise. See also tutorial below.

1.e Is the HoG descriptor invariant to overal lighting, *i.e.*, is the HoG descriptor of $I = 2 * I_A$ equal to the HoG descriptor of I_A?

Answer: No. It is invariant to additive color ($I = I_A + c$), but not to scaling. For more details see the nice hands-on tutorial on: https://www.learnopencv.com/histogram-of-oriented-gradients/

2 Bag-of-Words

2.a Describe the basic steps of the Bag-of-Visual Words model?

Answer: Offline stage: compute BoW vocabulary (see 2.e). Then, for each image: (1) Sample patches; (2) Compute per patch the descriptor; (3) Assign to closest word in the vocabulary.

2.b What is the difference between dense sampling en interest point sampling?

Answer: Finding interestpoints on salient parts of the image (eg using Harris) versus a dense (multi-scale) sampling grid. Advantage of interest points: focusses on salient parts of the image/object; advantage of the dense sampling strategy: finding high quality/good interest points is difficult, dense sampling ensures you have an even number of sampled patches per image, on any place in the image.

2.c Assume we have an image retrieval system, with 5000 images, 100 query images, and we use a BoW representation with 10K words, using SIFT descriptors. We observe that a BoW with *dense sampled* patches outperform BoW with *interest points* patches, when using precision@10 as evaluation measure. What could be the reason?

Answer: Only a very few images, so possibly the relevant images do not have enough interest points to get stable descriptors.

2.d Now we increase the dataset to 5M images, and interest points perform better. What could be the reason?

Answer: Precision is more important than recall (in this evaluation measure), so with the much larger dataset, the interest points could be able to find better matches (ie object only).

2.e Explain how k-Means clustering can be used to obtain a visual vocabulary.

Answer:

1. Sample a large set of patches from the train set (1M)
2. For each patch get the descriptor
3. Run K-Means over this set of descriptors
4. The resulting means are your *visual words*
5. The value of k is a hyperparameter
6. In practice it works better to compare a few random initialisation with only one or two iterations of k-means, than have a single random initialisation and run to convergence

3 Retrieval

3.a For retrieval, each image is described with 1000 interest points, each interest point is 128 dimensional. When finding matches between 2 images, how many computations are required?

Answer:

– compare each intrest point in image 1 with each interest point in image 2
– each comparison takes 128 computations
– total of 1000 * 1000 * 128 = 128M

3.b Comparing 1M interest points, takes 1 second. How long does it take to compare an image with a dataset containing 1M images?

Answer:

– Note here then granularity is ”comparing interest points”
– So, comparing two images takes 1 second (see question above).
– Answer: 1M seconds (not taking into account the sorting).

3.c After retrieving results with BoW, we use geometrical verification to rerank the top 100 images. However, our evaluation shows no difference in precision and recall measured at k=100. Why?

Answer: We evaluate the set of 100 images for retrieval and recall, re-ordering these would not change precision and recall at 100. It could improve, *eg* precision and recall at 10.

3.d Explain why accuracy is not a good metric

Answer: Accuracy is defined as the average number of ”correct assignments”; Both relevant and non-relevant images are taken into account. Given that for retrieval most images are not relevant, always returning ”no image at all” give a very high accuracy, but not a sensible retrieval system.

3.e Compute Average Precision for the following relevance ranking: $[R, N, R, R, N]$, where R denotes relevant, and N not-relevant

Answer:

$$\text{AP} = \frac{1}{R} \sum_r P(r) * R(r), \tag{1}$$

$$= \frac{1}{3} \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{4} \right) \approx 0.80, \tag{2}$$

where R is total number of relevant documents, $P(r)$ is the precision at rank r, and $R(r)$ is the relevance of the document at rank r.

4 Classification & Object Detection

4.a Explain how to train a ”cat” vs ”non-cat” classifier using linear classification (ie SVMs)

Answer:

– collect a large dataset of annotated images
– split into train, validation, and test set
– compute BoW vocabulary over train and validation
– compute BoW representation of all images (train, val and test)
– use train set to train your favourite classifier, use val set to select hyperparameters (number of words, regularisation, etc).
– evaluate the performance of your classifier on the test set.

4.b Compute the number of box evaluations for a single image in a multi-class object detection problem

Answer: The total number of evaluation is: #locations $\times$ #aspect ratios $\times$ #scales $\times$ #classes

4.c Explain the idea of Selective Search?

Answer: Find a (small) set of class agnostic object bounding boxes

4.d How does Selective Search increases the variations?

Answer: At least by using hierarchical clustering and using different color spaces