Object recognition and computer vision, Master MVA, 2017/2018

# Assignment 2: Image classification

Adil Rhoulam

arhoulam@ens-paris-saclay.fr

**QA1: Why is the spatial tiling used in the histogram image representation?**

(i) It is important to use spatial tiling in the histogram image representation to adds spatial information to BoVW in order to increase the performance rate. If we don't use the spatial tiling, even if we make BoVW flexible to viewpoint changes, the lack of spatial information not taking into account can lead to poor performances at places where the spatial information is a discriminant factor.

**QB1: Show the ranked training images in your report.**

(i) The ranked training images after training the classifier are shown, with their scores, in Figure 1.

**QB2: In your report, show relevant patches for the three most relevant visual words (in three separate figures) for the top ranked training image. Are the most relevant visual words on the airplane or also appear on background?**

(i) In the figures that display the patches for the three most relevant visual words for the top ranked training image, we can notice that the most relevant visual words are not necesseraly on the airplane like in a) Figure 2, the most relevant visual word is not on the background.

**QC1: Why is the bias term not needed for the image ranking?**

(i) the bias term not needed for the image ranking because it does the same translation (t=bias) to all the scores of the images, then it doesn't change the ranking of the images, but only the values of the scores which will be translated.

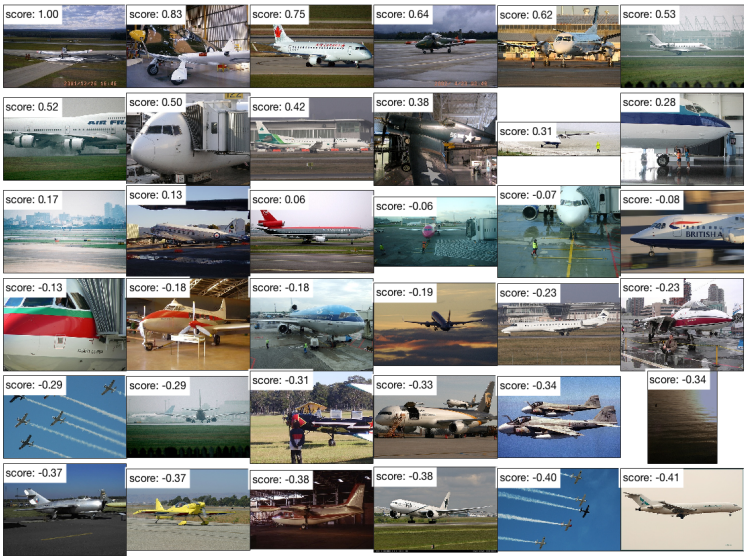
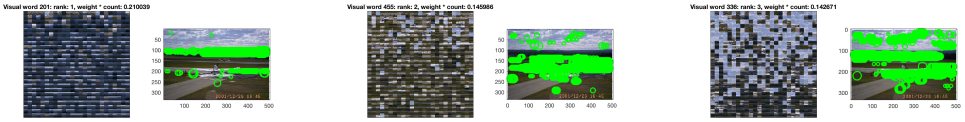



Figure 1: Ranked training images



(a) First relevant visual word (b) Second relevant visual word (c) Third relevant visual word

Figure 2: Patches for the three most relevant visual words for the top ranked training image

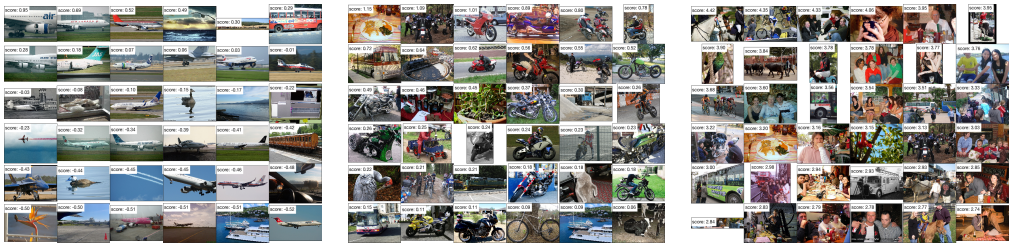
**QD1: In your report, show the top ranked images, precision-recall curves and APs for the test data of all the three classes (aeroplanes, motorbikes, and persons). Does the AP performance for the different classes match your expectations based on the variation of the class images?**

(i) The ranked test images are shown, with their scores, in Figure 3 for different datasets.

(ii) The Precision-recall curves for the test data of all the three classe are shown in Figure 4.

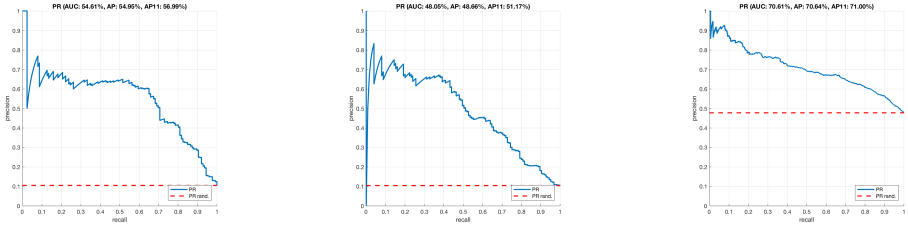
(iii) The APs for different classes for the test data are:
- Test AP: 0.55 airplane
- Test AP: 0.71 person
- Test AP: 0.48 motorbike

(iv) The AP performances were expected, the large AP value goes for the Person class



(a) Airplane classe (b) Motorbike class (c) Person class

Figure 3: Ranked test images with their scores for the three classes




(a) Airplane classe (b) Motorbike class (c) Person class

Figure 4: The Precision-recall curves for the test data of all the three classes

because, its training set (positive images) contains more data than the other classes, so the model is well trained. For the two other classes, the AP value is not great, this is due to the lack of positive images comparing with negative images given to the classifier to train.

**QD2: For the motorbike class, give the rank of the first false positive image. What point on the precision-recall curve corresponds to this first false positive image? Give in your report the value of precision and recall for that point on the precision-recall curve.**

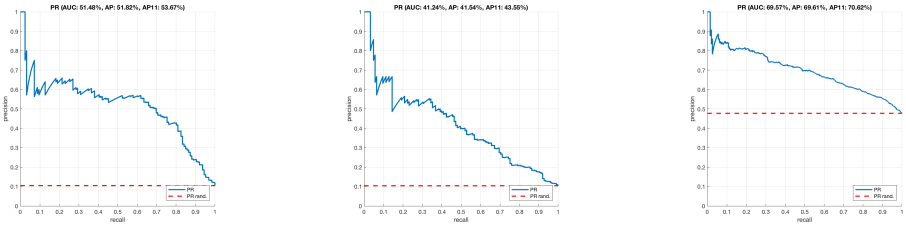
(i) False positive images is the set of negative images that were classified positive by the classifier, so we are seeking for the first ranked images that has a positive score and its label is -1, the first image has a positive score 1.05 and its label is given by testLabels(perm(1)) which return -1, so it's the first ranked false positive image shown in Figure 5.

(ii) We get the value of the Precision Recall of the point by the two commands: $[RECALL, PRECISION] = vl\_pr(testLabels, testScores)$; $RECALL(perm(1))$ and $PRECISION(perm(1))$;
We get as output: Precision = 0.140811 and Recall = 0.944000 are the values of the Precision-Recall curve in the first ranked false positive image point.


Figure 5: First false positive ranked image

**QE1: Include in your report precision recall-curves and APs, and compare the test performance to the spatially tiled representation in stage D. How is the performance changing? Why?**

(i) The Precision-recall curves for the test data when removing the spatial tiling of all the three classes are shown in Figure 6.



(a) Airplane classe (b) Motorbike class (c) Person class

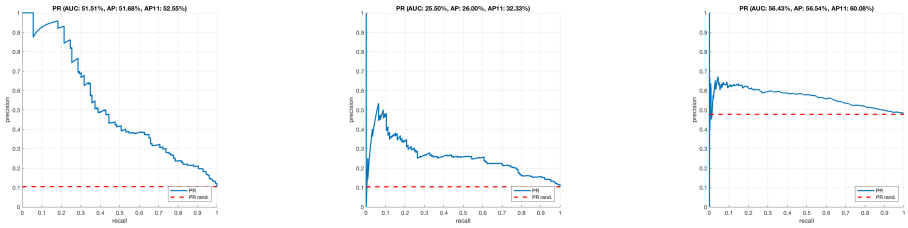
Figure 6: The Precision-recall curves for the test data of all the three classes after removing the spatial tiling

(ii) The APs for different classes for the test data are:
- Test AP: 0.51 airplane
- Test AP: 0.70 person
- Test AP: 0.41 motorbike

(iii) The AP performances have decreased after removing the spatial tiling, this is because, as said in the first question, it adds spatial information to BoVW and then the performance increases.

**QE2: Modify exercise1.m to use L1 normalization and no normalization and measure the performance change.**

(i) The Precision-recall curves of test data, when applying the L1 norm to normalize the histograms, for all classes are shown in Figure 7.



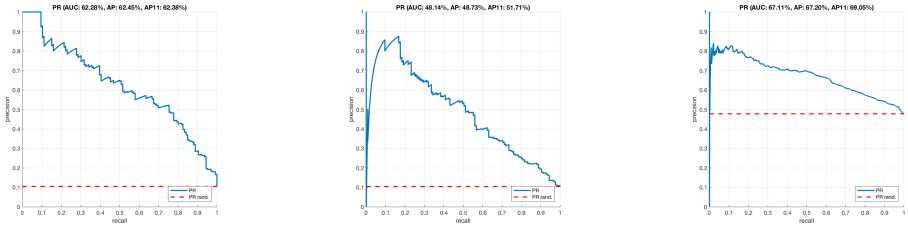

(a) Airplane classe (b) Motorbike class (c) Person class

Figure 7: The Precision-recall curves of test data with L1 histograms normalization

(ii) The APs for different classes for the test data are:
- Test AP: 0.51 airplane
- Test AP: 0.56 person
- Test AP: 0.25 motorbike

The AP performances have decreased for all the datasets when normalizing the histograms with L1 norm instead of L2.

(iii) The Precision-recall curves of test data, without normalizing the histograms, for all classes are shown in Figure 8.




(a) Airplane classe (b) Motorbike class (c) Person class

Figure 8: The Precision-recall curves of test data without histogram normalization@

(iv) The APs for different classes for the test data are:
- Test AP: 0.62 airplane
- Test AP: 0.67 person
- Test AP: 0.48 motorbike

The AP performances have decreased for Person datasets, increased for Airplane dataset and stayed almost the same for the motorbike datasets when the histograms aren't normalized.

**QE3: What can you say about the self-similarity, K(h,h), of a BoVW histogram h that is L2 normalized? Hint: Compare K(h,h) to the similarity, K(h,h'), of two different L2 normalized BoVW histograms h and h'. Can you say the same for unnormalized or L1 normalized histograms?**

(i) If we denote g, g' the histogram correspondent to h, h' without L2 normalization (i.e. h = $\frac{g}{||g||_2}$ and h' = $\frac{g'}{||g'||_2}$)
K(h,h') = $\sum_{k=1}^{d} \frac{g_k . g'_k}{||g||_2 . ||g'||_2}$
then K(h,h) = $\sum_{k=1}^{d} \frac{g_k . g_k}{||g||_2 . ||g||_2} = \sum_{k=1}^{d} \frac{g_k^2}{||g||_2^2} = \frac{1}{||g||_2^2} \cdot \sum_{k=1}^{d} g_k^2 = \frac{||g||_2^2}{||g||_2^2} = 1$
Finally K(h,h) = 1 for a BoVW histogram h that is L2 normalized.

(ii) We can't say the same thing for L1 normalized histograms because we get: K(h,h) = $\frac{||g||_2^2}{||g||_1^2} \neq 1$ in general.

**QE4: Do you see a relation between the classification performance and L2 normalization?**

(i) With L2 normalization, the performance of classification increases. This is because when using a linear kernel, as seen in the previous question, the vectors are normalized w.r.t to the linear kernel (k(h,h)=1) while it's not for the L1 norm which could lead to inconsistent similarity measures between different histograms since vectors hasn't the same kernel norm (k(h,h) $\neq$ 1 ).

**QF1: Based on the rule of thumb introduced above, how should the BoVW histograms h and h' be normalized? Should you apply this normalization before or after taking the square root?**

(i) Based on the rule of thumb introduced in the previous question, the BoVW histograms h and h' should be normalized after taking the square root, so that the vectors remain normalized w.r.t to the Hellinger kernel. In fact if we take normalize after taking the square root we have:
$\frac{\sqrt{g}}{||\sqrt{g}||_2}$ and $\frac{\sqrt{g'}}{||\sqrt{g'}||_2}$ are the histograms in the new feature space.
If h and h ' are non normalized histograms:
K(h,h') = $\sum_{k=1}^{d} h_k . h'_k = \sum_{k=1}^{d} \frac{\sqrt{g}_k}{||\sqrt{g}||_2} . \frac{\sqrt{g'}_k}{||\sqrt{g'}||_2}$
We get K(h,h) = 1

**QF2: Why is this procedure equivalent to using the Hellinger kernel in the SVM classifier?**

(i) This procedure is equivalent to using the Hellinger kernel in the SVM classifier because we sum the product of the square of the two histograms values :
(K(g,g') =C(g,g') . $\sum_{k=1}^{d} \sqrt{g}_k \cdot \sqrt{g'}_k$ ) where C(g,g') is the normalization constant that we have used after taking the square root to preserve the linearity of the classifier, so the term containing the sum is exactly the Hellinger kernel.

**QF3: Why is it an advantage to keep the classifier linear, rather than using a non-linear kernel?**

(i) The advantages to keep the classifier linear, rather than using a non-linear kernel, are:
- The linear classifiers are easy and fast to train and test (speed classification), while the non-linear classifiers have a high runtime complexity and need much more time to do so.
- The linear classifiers need less spatial memory requirements than the non linear one thanks to the compactness of their decision functions.

**QF4: Try the other histogram normalization options and check that your choice yields optimal performance. Summarize your finding in the report (include only mAP results, no need to include the full precision-recall curves).**

(i) With L2 normalization using the Hellinger kernel in the SVM classifier we get the following APs results:
- Test AP: 0.71 airplane
- Test AP: 0.77 person
- Test AP: 0.63 motorbike
- mAP = (0.71+0.77+0.63 )/3 = 0.70

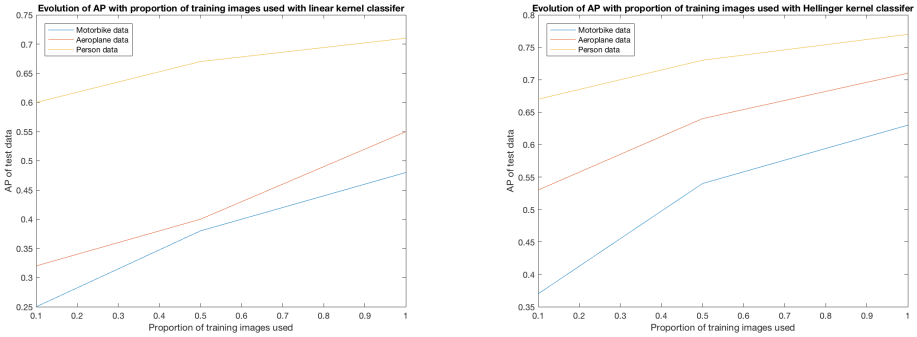
With L1 normalization using the Hellinger kernel in the SVM classifier we get the following APs results:
- Test AP: 0.55 airplane
- Test AP: 0.58 person
- Test AP: 0.21 motorbike
- mAP = (0.55+0.21+0.58)/3 = 0.44

Without normalizing using the Hellinger kernel in the SVM classifier we get the following APs results:
- Test AP: 0.65 airplane
- Test AP: 0.69 person
- Test AP: 0.55 motorbike
- mAP = (0.65+0.69+0.55)/3 = 0.63

Finally, we can conclude that the choice of L2 norm yields the optimal performances.

**QG1: Report and compare performance you get with the linear kernel and with the Hellinger kernel for the different classes and proportions of training images (10% ,50% and 100%). You don?t have to report the precision-recall curves, just APs are sufficient. Plot the APs for one class into a graph, with AP on the y-axis and the proportion of training images on the x-axis. You can use the matlab function plot Plot three curves (one curve for each class) into one figure. Produce two figures, one for the linear kernel and one for the Hellinger kernel. Make sure to properly label axis (use functions xlabel and ylabel), show each curve in a different color, and have a legend (function legend) in each figure. Show the two figures in your report.**

(i) The two plots are shown in Figure 9.




(a) Using Linear Kernel (b) Using Hellinger kernel

Figure 9: Evolution of AP with different proportions of training images

**QG2: By analyzing the two figures, do you think the performance has 'saturated' if all the training images are used, or would adding more training images give an improvement?**

(i) By analyzing the two figures, we can see that the performances increases when adding more training images. So for the Motorbike and Airplane data it's worth to add training data, however it's like the performance converges to a certain value for the person data so when testing for the Hellinger kernel 90% of training data we get 0.61 AP and for 100% data we get 0.63 AP, so adding data may increases the performances but not remarkably ( We need to add a lot of data to notice a remarkable improvement of AP).

**QP2.1: For the horse class, report the precision at rank-36 for 5 and 10 training images. Show the training images you used. Did the performance of the classifier improve when 10 images were used?**

(i) Correctly retrieved in the top 36: 4 , then Precision at rank 36 is = $\frac{4}{36}$= 0.11 for 5 positive training images (horses images).
Correctly retrieved in the top 36: 8 , then Precision at rank 36 is = $\frac{8}{36}$= 0.22 for 10 positive training images (horses images). The performance of the classifier has improved when adding 5 more horse images to the training set.

(ii) The training images used are shown in Figure 10.

**QP2.2: What is the best performance (measured by precision at rank-36) you were able to achieve for the horse and the car class? How many training images did you use? For each of the two classes, show examples of your training images, show the top ranked 36 images, and report the precision at rank-36. Compare the difficulty of of retrieving horses and cars.**

(i) For the horses data: Correctly retrieved in the top 36: 13 is the best performance (Precision at rank 36 is = 0.36) I achieved for Horses class using 26 training images (I used till 48 images but got less performance)

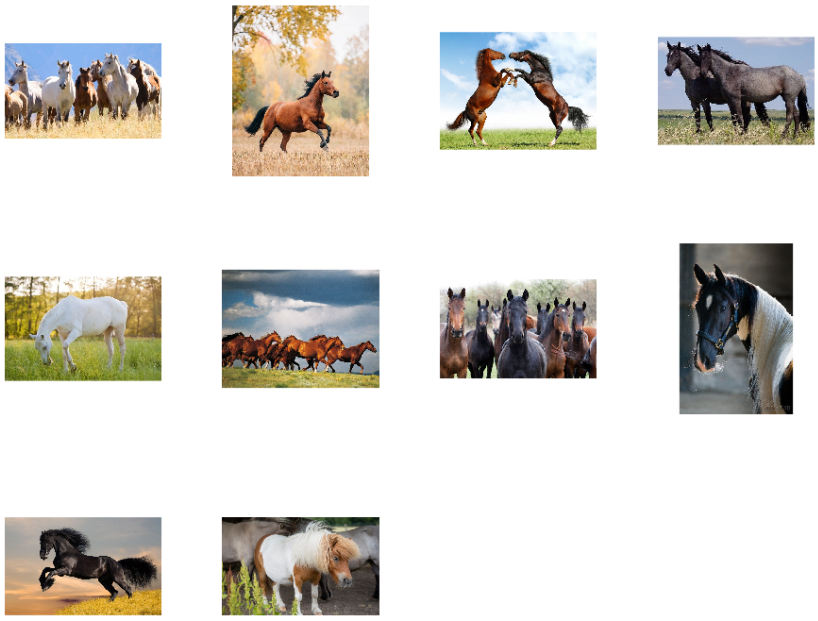


Figure 10: Horses images used for training

(ii) For the Cars data: Correctly retrieved in the top 36: 31 for 10 positive training images then Precision at rank 36 is $= \frac{31}{36} = 0.86$ for 10 positive training images that are shown in Figure 11 .
The best performance I get is: Correctly retrieved in the top 36: 35 (Precision at rank 36 is 0.97) for 21 positive training images.

(iii) We conclude that the horse's class is difficult to retrieve than the car's class, this is because we can find a lot of cars images with a white or unicolor background which help the classifier to learn rapidly, while for the horses class it's difficult to find images with a white or unicolor background.

**QH1: Compare the dimension of VLAD and BoVW vectors for a given value of K. What should be the relation of the K in VLAD to the K in BoVW in order to obtain descriptors of the same dimension? You can ignore tiling.**

(i) In the BoVW approach, after calculating the visual words with the K-means algorithm, each vector in the SIFT descriptor is mapped into an integer index between 1 and K based on the Nearest Neighborhood visual word, so the dimension of the BoVW vector is Kx1. In the VLAD approach, the dimension of the vector is KxD for each visual codeword, we assign to it the sum of the residuals lying in it cluster.

(ii) The relation of the K in VLAD to the K in BoVW in order to obtain descriptors of

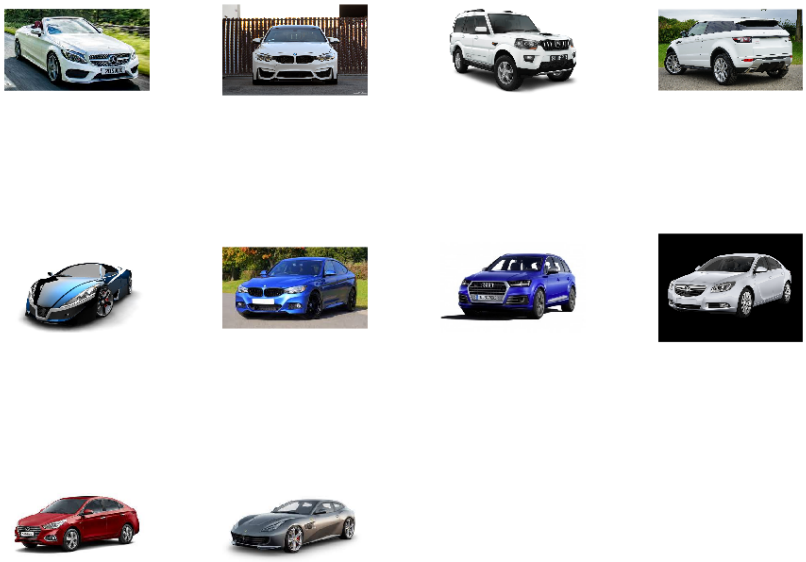


Figure 11: Cars images used for training

the same dimension is: K(BoVW)=D*K(VLAD).

**QH2: Replace the encoding used in exercise1 with the VLAD encoding, and repeat the classification experiments for the three classes of Part I (Both linear and Hellinger kernel). How do the results compare to the BoVW encoding? Report mAP results in a table. No need to report all precision-recall curves.**

(i) See Table1 for the informations requested. Note that: LK stands for Linear Kernel and HK stands for Hellinger Kernel.

(i) We can notice that the results have improved with the linear Kernel and the perfor-

| Class | VLAD + LK | VLAD + HK | BoVW + LK | BoVW + HK |
|---|---|---|---|---|
| Airplane AP | 0.75 | 0.07 | 0.55 | 0.71 |
| Motorbike AP | 0.69 | 0.07 | 0.48 | 0.63 |
| Person AP | 0.76 | 0.55 | 0.71 | 0.77 |
| mAP | 0.73 | 0.23 | 0.58 | 0.70 |

Table 1: Comparison between BoVW and VLAD performances

| Class | VLAD + LK | VLAD + HK | BoVW + LK | BoVW + HK | FV + LK | FV + HK |
|---|---|---|---|---|---|---|
| Airplane AP | 0.75 | 0.07 | 0.55 | 0.71 | 0.70 | 0.07 |
| Motorbike AP | 0.69 | 0.07 | 0.48 | 0.63 | 0.73 | 0.10 |
| Person AP | 0.76 | 0.55 | 0.71 | 0.77 | 0.77 | 0.52 |
| mAP | 0.73 | 0.23 | 0.58 | 0.70 | 0.73 | 0.23 |

Table 2: Comparison between BoVW, VLAD and FV performances

mance has decreased when using the non linear Kernel (Hellinger).

**QI1: Replace the encoding used in exercise1 with the FV encoding, and repeat the classification experiments for the three classes of Part I. Report the results in the same table as QH2 so that you can see the performance of the three encoding methods side by side.**

(i) See Table2 for the informations requested. Note that: LK stands for Linear Kernel and HK stands for Hellinger Kernel.

(i) We can notice that the mAP results given by VLAD and FV are the same and then have the same mean performances: Gives better results than BoVW for a LK and less performances than BoVW for HK.

**QI2: What are the advantages or disadvantages of FV compared to VLAD in terms of computation time and storage/memory footprint - especially for a large number (hundreds of millions) of images.**

(i) The VLAD is similar to Fisher vectors but it doesn't store second order (variance) information about the features and it uses K-means instead of GMM to generate the feature vocabulary, we now that GMM is more computationally expensive than K-means, then especially for large number of images FV is computationally and memory expensive than VLAD because it uses GMM and it's store the Variances of the features in addition to the means of Features. The advantages of FV over VLAD is that FV is more efficient and uses a soft assignment for the codewords (GMM).