TBMI26 – Computer Assignment Reports Boosting

Deadline – March 14 2021

Author/-s: Wuhao Wang (wuhwa469)

In order to pass the assignment you will need to answer the following questions and upload the document to LISAM. Please upload the document in PDF format. **You will also need to upload all code in .m-file format**. I will correct the reports continuously so feel free to send them as soon as possible. If you meet the deadline you will have the lab part of the course reported in LADOK together with the exam. If not, you'll get the lab part reported during the re-exam period.

1. **Plot how the classification accuracy on training data and test data depend on the number of Iak classifiers (in the same plot). Be sure to include the number of training data (non-faces + faces), test-data (non-faces + faces), and the number of Haar-Features.**

2. **How many Iak classifiers did you use when training? How many of them did you use for the final strong classifier? Motivate your choices.**

   I use 50 Iak classifiers in training and all of them are used in the final strong classifier. Since during the training stage, I apply a strategy that no sample can to too 'important'. In other words, I set a upper limitation for single sample weight .

$$\boldsymbol{Sample\ weight} \leq \frac{\mathbf{1}}{\boldsymbol{Number\ of\ training\ sample}/\mathbf{10}}$$

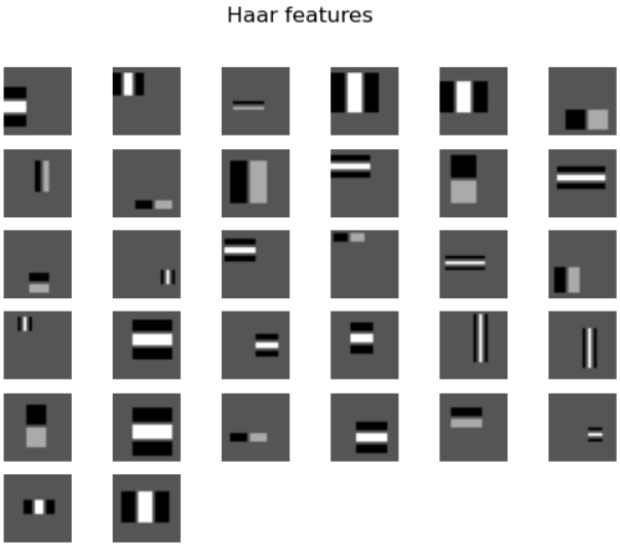
   So, no classifier would overfit. I also tried to drop some classifiers at the beginning of inference stage but there was no improvement.

3. **What is the accuracy on the training data and test data after applying the optimized strong classifier? Discuss your choice of hyperparameters and how they influence the accuracies.**

   The accuracy is **0.965** on training data and **0.93465** on test data.
   Choice of hyper parameters:
   number of Haar-Features: **60** , number of Training Images:**2000,**
   number of weak Classifier:**50**, sample weight limitation:**0.005**

   More **Haar-features** will provide more description of the data, thus helping improve the accuracy. More **training images** will make the training distribution be more closed to real data and provide more cut-off value, thus helping the stump knows better about how to make decision. More **Iak Classifiers** can correct the previous mistakes but may also lead to overfit. LoIr **sample Iight limitation** will easily lead to underfitting, while too large scale will make this parameter useless.

4. **Plot the Haar-features selected by your classifier (one for each weak classifier). If you have many Iak classifiers, select some representative subset. Can you think of why they would be useful for classifying faces?**



   I can see that the weakest classifiers focus on the center of the picture. And in the plot, I can find that the color of eyes, eyebrow, mouth or lips are different with the surrounding skin. Meanwhile, because of the face structure, if I draw a vertical line alone the nose, this line could be the 'protruding part' of the face which make the brightness here be different with the rest of the face.

5. **Plot some of the misclassified faces and non-faces that seem hard to classify correctly. Why do you think they are difficult to classify?**

In the misclassified-nonfaces images, many of them are too dark or too bright which make it difficult to extract 'human face' features. And some people are wearing glasses ,this will make the pixel value distribution different from original face. So, I think glasses might be a factor.

In the misclassified-faces images, in the line 3 , the third and fourth images are hard to identify even with our eyes. For other images, some of them have the contour being similar to human faces, some of them have a vertical line which are similar to nose on the pixel level.

6. **Are your results reasonable? Can you think of any way to improve the results?**
   I think the results are reasonable.
   The way to improve:
   1 train with more Haar-Features and classifiers.
   2 train multiple strong classifiers and make them work together.

7. **Can I expect perfect results? Motivate your answer.**

   The perfect results mean that the strong classifier should be completely correct on both training and test data. Under most circumstances this is impossible.

   Training data and test data can hardly have totally same data distribution which means the knowledge learned from training data might be biased. Besides, the training data may contain some outliers which will lead the model to be away from the 'truth'.