Assignment 1: Raw Data to Feature Space

Due Date: September 17th, 2022

In the emerging field of data science, especially in big data and machine learning, there are four essential objectives, and they are: (a) understanding of data; (b) understanding of machine learning; (c) understanding of systems, and (d) understanding of scalability and complexity. In addition, it is important that the data analysts and data scientists to learn new big data systems (e.g., Hadoop, Spark) and programming languages (e.g., R, Python, Scala). The report writing and presentation skills are also equally important; therefore, you must prepare your documents using Latex with well-known IEEE, ACM, Springer, or Elsevier formats for the assignments.

The main goal of this assignment is to extract features from a set of given data (generally from raw data) and construct a feature space or feature spaces (along with subspace as appropriate) to develop and train machine learning models. It helps test your knowledge in the first objective; that is, knowing and transforming your data (Chapters 1, 2, and 3) with feature extraction. It helps you understand the characteristics of data such that correct computational and ML techniques and technologies can be selected to process and analyze the data.

Task 1: Build your programming environment!

- https://docs.anaconda.com/anaconda/ (Download and install a Python environment). Python is used as programming language to explain the rest of this assignment, but you may use any suitable language
- https://docs.anaconda.com/anaconda/user-guide/tasks/integration/spyder/ (Install the Spyder IDE)
- https://jupyter.readthedocs.io/en/latest/install.html (Install the Jupyter Notebook, if needed)
- Download and install OpenCV for using computer vision tools in this assignment

➢ *Submit (a report) screenshots of your programming environment along with the commands that you used successfully to install the components and configurations of your final programming environment*

Task 2: Generate a fruits and vegetables image dataset!

- Use the fruits image that I provided with this assignment first
- Download these images from Canvas and store them in a folder
- Make sure you extract gray scale images from these color images for this assignment
- Select three types of fruits to perform an initial study (e.g., lime, orange, and apple). You could also use your cell phone to take pictures of fruits and vegetables and generate your own fruits and vegetable image datasets. You must justify the selection of your images with computational reasoning – chapter 3 from the textbook should help you find computational justifications. Name your images as image0, image1, and image2 and use them to carry out this assignment.

➢ *Please store these selected images in a folder and submit it with a report (a pdf document). Use a suitable subheading for this section in the report. Only the selected images must be included in the report.*

Task 3: Write a simple code to read your selected images and display them on the programming environment!

- Select and add suitable Python and OpenCV libraries to your code to perform this task
- Write a code to read these color images and display their R, G, and B channel images
- Convert these color images to grayscale and display them while printing their dimensions

➢ *Add the code and results to the document is being generated - use a suitable subheading*

Task 4: Resize the images to reduce their dimensions!

- Write a function (Python) to resize (reduce) the grayscale images such that the output dimensions are divisible by 12 without changing their original aspect ratios significantly
- You must parametrize the function appropriately to generalize the reduction process
- Use this parametric function to reduce the size of the grayscale images to have height of 264 pixels – approximately maintain the aspect ratio for the width, but it must be divisible by 12
- ➢ *Add the code and the results to the document - use a suitable subheading*

Task 5: Generate block-feature vectors!

- Write a code to divide each image into blocks of 12x12 pixels and transform them to vectors of size 144
- Assign a label to each feature vector with 0, 1, and 2 for the first, second, and third images, respectively
- Generate a spreadsheet by storing a feature vector per row in the spreadsheet for each image
- A visual illustration is provided in Figure 1 for generating block-feature vectors using 8x8 pixel blocks, but you need to perform the same tasks for 12x12 pixel blocks. You need to perform this operation on the entire image spatially to generate feature vectors of dimension 144
- ➢ *Add the code and the results to the document - use a suitable subheading*

Task 6: Generate sliding block-feature vectors!

- Write a code to divide an image into sliding blocks of 12x12 pixels and transform them to feature vectors
- Assign a label to each feature vector with 0, 1, and 2 for the first, second, and third images, respectively
- Generate a spreadsheet by storing a feature vector per row in the spreadsheet for each image
- A visual illustration is shown in Figure 2 for generating sliding block-feature vectors using 8x8 pixel blocks, but you need to perform the same tasks for 12x12 pixel blocks
- ➢ *Add the code and the results to the document - use a suitable subheading*

Task 7: Derive statistical descriptors!

- Extract statistical information (e.g., number of observations, dimension of the data, mean of each feature, etc.) from these datasets. Also, present visual representations (e.g., histogram, scatter plot, etc.) of the data.
- Answer the following questions -- Is the dataset imbalanced, inaccurate or incomplete? Is it a trivial data or possibly a big data? Does it have scalability problem? Are they high dimensional? Do you need to standardize? Do you need to normalize? How do they affect the data characteristics?
- You must think about the above questions/problems and provide your explanation scientifically. You need to write programs to read the data and generate results to explain all of the above – since you need to show/justify.
- You can follow chapter 3 discussions to answer the above questions. This chapter provides required details based on the analysis of two sets of images: (a) carpet and hardwood floor and (b) Biltmore and PrismaColors
- ➢ *Add the code and the results to the document - use a suitable subheading*

Task 8: Construct a feature space!

- Merge the feature vectors in image0.csv and image1.csv to create a feature space for these images. Each feature and label columns must align vertically to generate the correct feature space for these image classes. Name the feature space file (spreadsheet) as image01.csv

- Similarly merge the feature vectors in image0.csv, image1.csv, and image2.csv to create a feature space for these images. Each feature and label columns must align vertically to generate the correct feature space for these three classes. Name the feature space file (spreadsheet) as image012.csv

- Randomize the placement of the feature vectors in the files image01.csv and image012.csv files. Note that you don't randomize the content of a feature vector, but the placement (rows) of the csv files. You can now see the labels are randomized and it will help the training of ML in the later assignment goals

  - *Add the code and results to the document is being generated - use a suitable subheading.*

Task 9: Display subspaces!

- Select two features and plot the two-dimensional feature space with labeling the observations (vectors) of the fruits or vegetables that you selected by using the spreadsheets that you generated
- Select three features and plot the three-dimensional feature space with labeling the observations (vectors) of the fruits or vegetables that you selected by using the spreadsheets that you generated
- You must generate separate plot to show two class labels (meaning two fruits or vegetables) and three class labels (meaning three fruits or vegetables)
- Discuss these figures and describe your observations in terms of their separable features

  - *Add the code and results to the document is being generated - use a suitable subheading.*

Task 10: Make appropriate changes to your Python code such that it can read any number of images from a folder that consists of many similar images, generate a feature space/s, and generate a spreadsheet/s for the feature spaces!

  - *Add the code and results to the document is being generated - use a suitable subheading.*

Task 11: Describe the effects of block size on the dimensionality of the feature space and the number of vectors in the domain. Also, describe how these effects may influence the classifier that divides the domain

  - *Add the discussion to the document that is being generated - use a suitable subheading.*

Task 12: Submit required documents!

- Prepare a Latex document using one of the IEEE, ACM, Springer, or Elsevier Latex formats. However, make sure you select two-column format
- Submit a zipped folder that consists of subfolders: (a) Latex subfolder that consists of all the necessary scripts and the pdf output (i.e., the report in two-column format); (b) Data subfolder that consists of all the images (both input and output), the spreadsheets with feature vectors and feature spaces that you have created based on the assigned tasks and the answers to all the questions; (c) a Code subfolder that consists of programs/modules that you developed to complete the task; and (d) a Screenshot subfolder that shows the programming environment that you created and the results that you obtained when you run your code
- This is an evidenced-based assessment; hence, it is your responsibility to submit all the required documents that show the completion of all the required tasks. Submit them as a zipped folder via Canvas.
- If you are in doubt or have questions send me an email: s_suthah@uncg.edu or visit during my virtual office hours. You can also ask questions and clear your doubts during zoom meetings. It is important that you do not make assumptions on assignment/test requirements based on the discussions with other students.

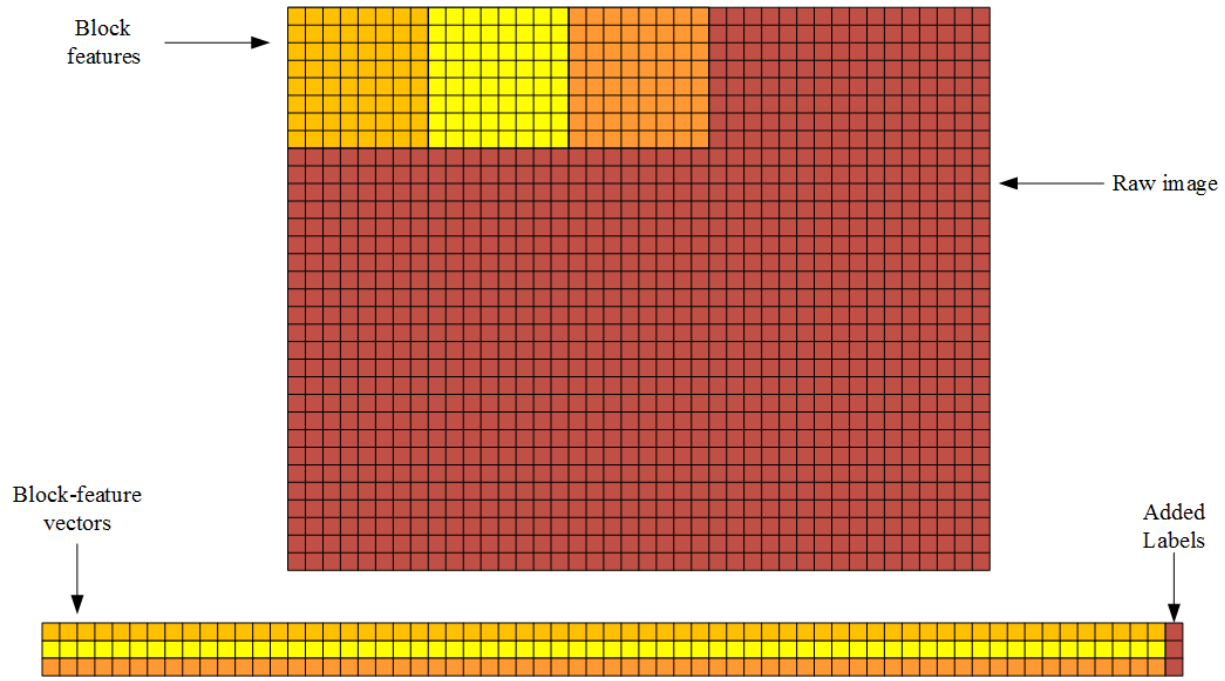

Figure 1: Generating block feature vectors from a raw image data – only three block-features are shown here, but you need to generate all the possible features from the entire image

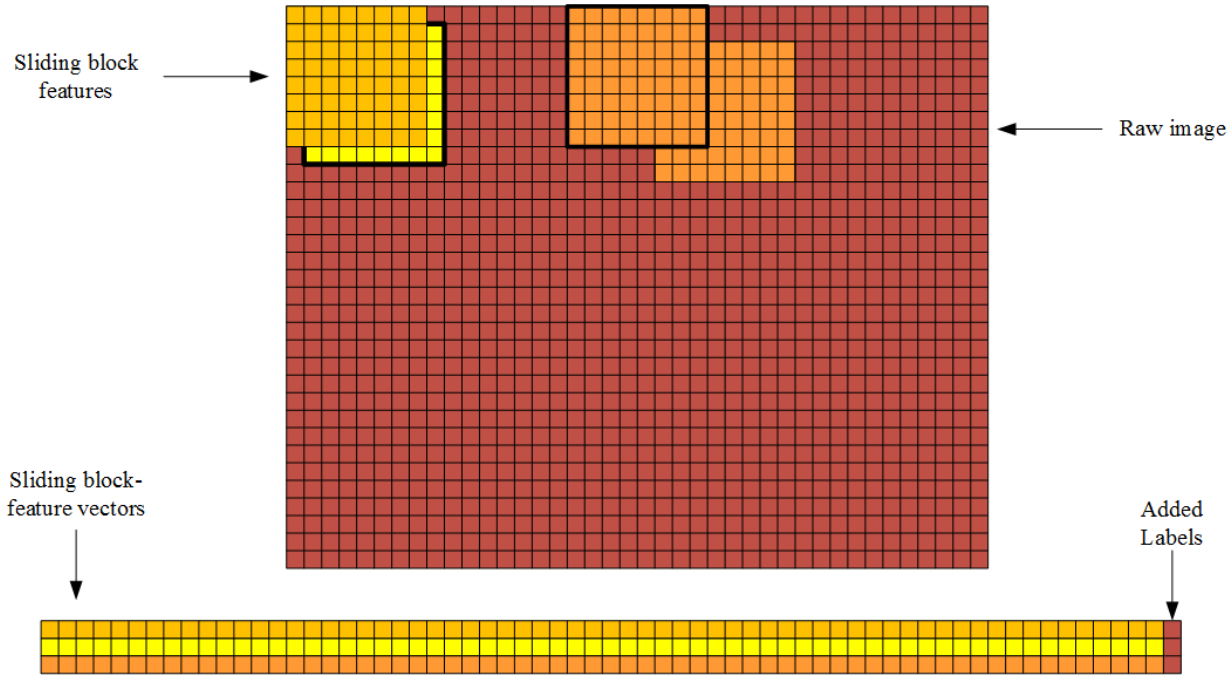

Figure 2: Generating sliding block feature vectors from a raw image data – only two sets of sliding block-features are shown here, but you need to generate all the possible features from the entire image