Assignment 3
Visualization using PCA

CMPE 481 Data Analysis and Visualization **Due: January 10th, 2021, 6am**

In this assignment, you are going to implement PCA method from scratch and use it to visualize the MNIST digit database in 2D. There are two approaches to implement the PCA method: 1) Using the covariance of the data, and 2) SVD approach. In this assignment, use the first approach. You can obtain the MNIST digit database from these web pages:

Original MNIST Dataset: http://yann.lecun.com/exdb/mnist/
CSV (Excel) Version: https://www.kaggle.com/oddrationale/mnist-in-csv

Your program should perform the following steps and your report should contain the results:

1. Plot 10 sample digit images per digit class as shown in Figure 1.
2. Using the MNIST training set: 1) plot the mean image, 2) eigenvectors, and 3) eigenvalues as shown in Figure 2 and Figure 3.
3. Using the MNIST test set, reduce the dimensionality of features to two for PCA visualization. Plot the digits in 2D as shown in Figure 4. Briefly explain your observations from the visualization, e.g., do you observe any patterns?
4. t-Distributed Stochastic Neighbor Embedding (t-SNE) is another popular visualization technique. Use t-SNE to visualize MNIST test set, as shown in Figure 5. You can use a toolbox for the t-SNE approach, e.g., scikit learn.
5. In one page, explain the fundamentals of the t-SNE approach and compare it to the PCA. What are the advantages and disadvantages of t-SNE approach compared the PCA?
6. Reconstruct two images using the PCA approach with different number of eigenvectors, e.g., i.e., first project to 2, 12, 22, 32, … 784 dimensions and then reconstruct the image. The first image should be taken from the MNIST test set, and the second image should be a non-digit (Be creative!). See Figure 6 as an example. For each of the images, provide the explained variance ratio as a percentage to achieve a good reconstruction using least number of eigenvectors as possible, e.g., it is possible to obtain a nice reconstruction with an explained variance ratio of X per cent.

Bonus [30 pts] Replicate all these steps with human faces.

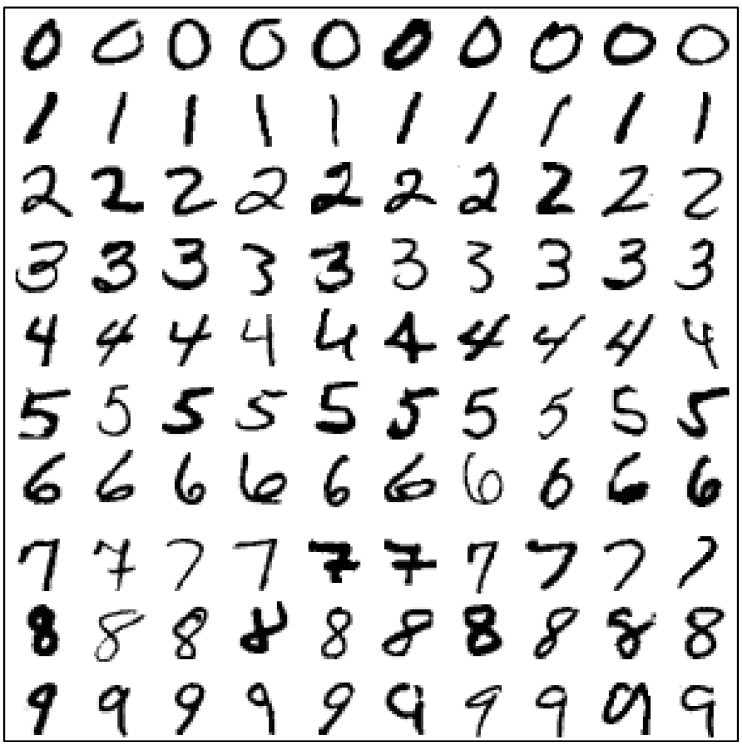

Figure 1. Sample digit images from the MNIST database.

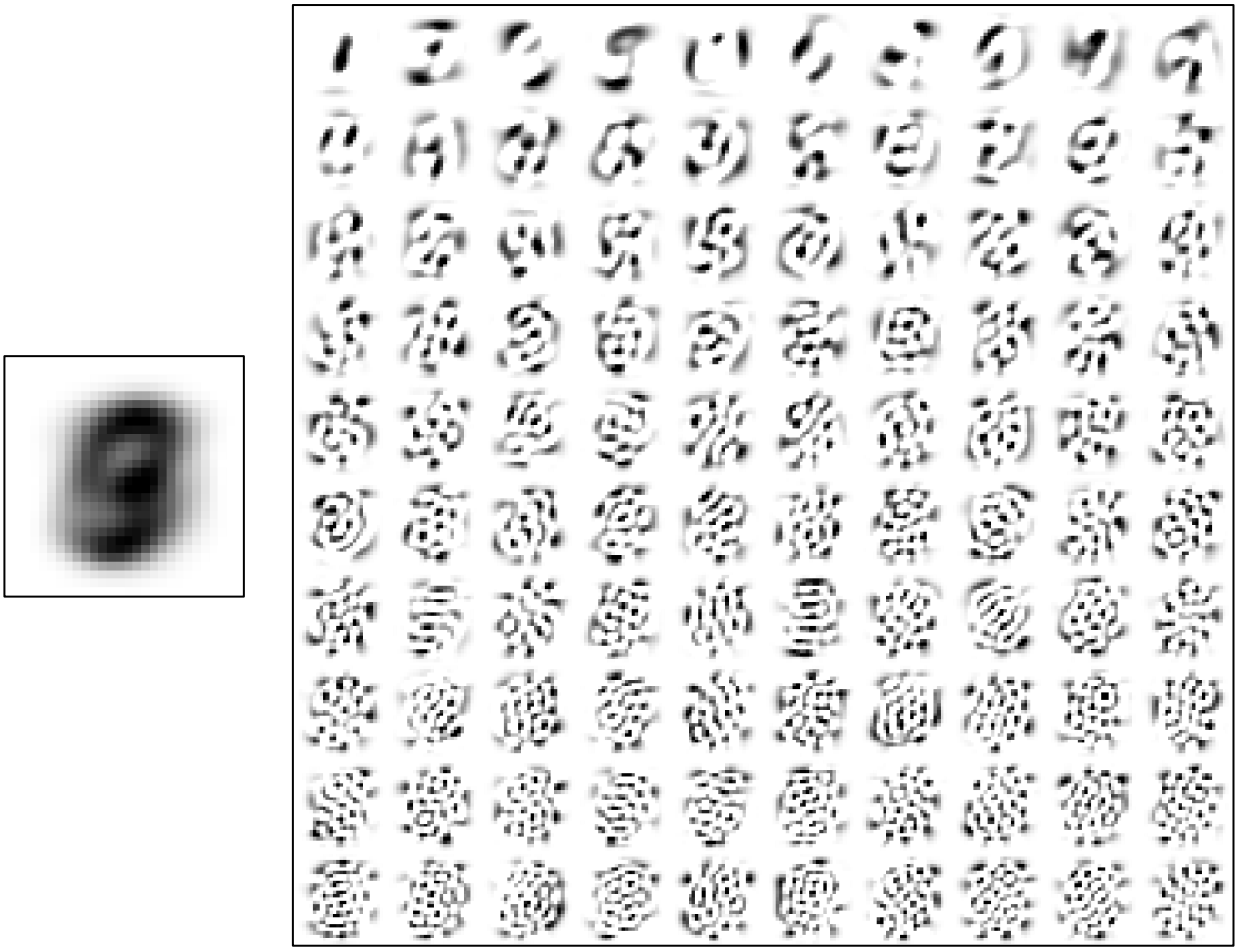

Figure 2. Left: Mean digit image. Right: Largest 100 eigenvectors.

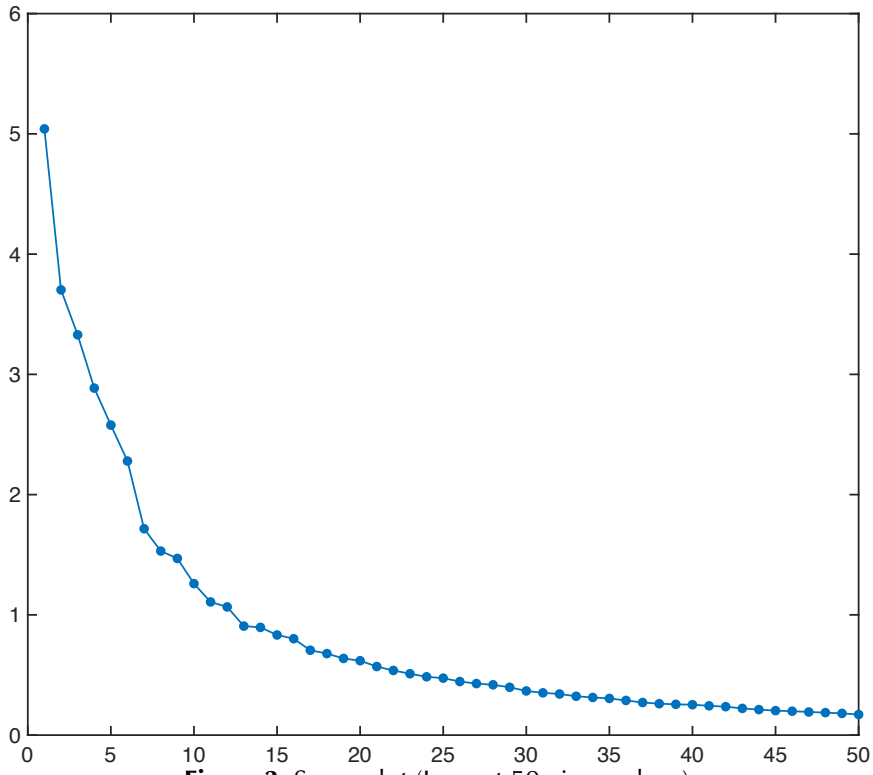

Figure 3. Scree plot (Largest 50 eigenvalues).

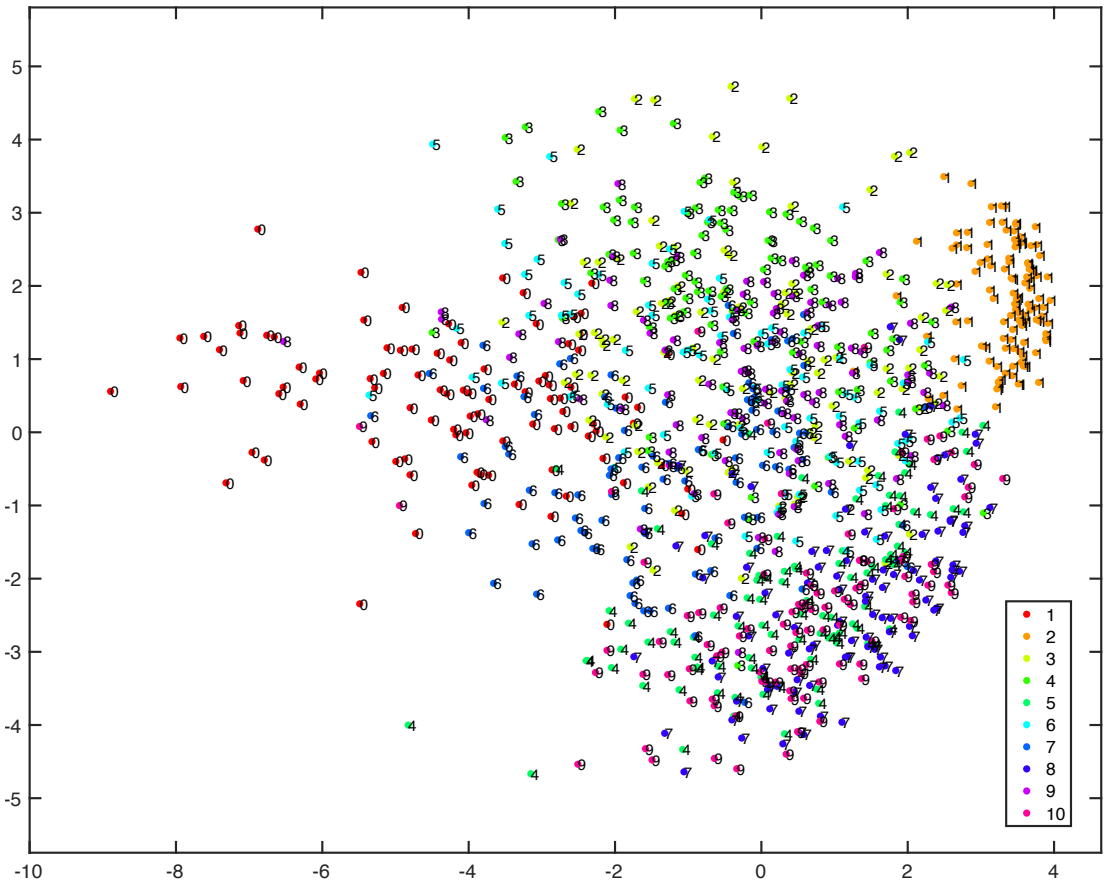

Figure 4. 2D PCA visualization of a subset of the MNIST test set.

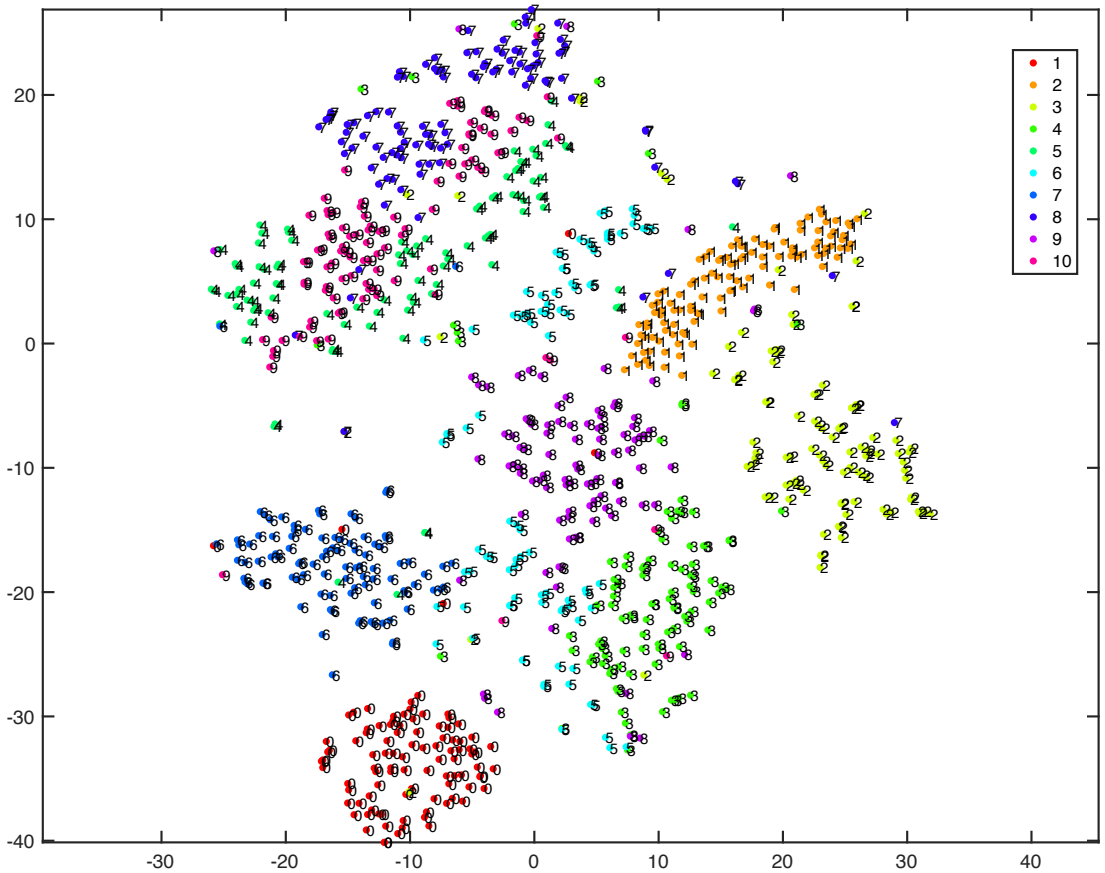

Figure 5. 2D t-SNE visualization of a subset of the MNIST test set.

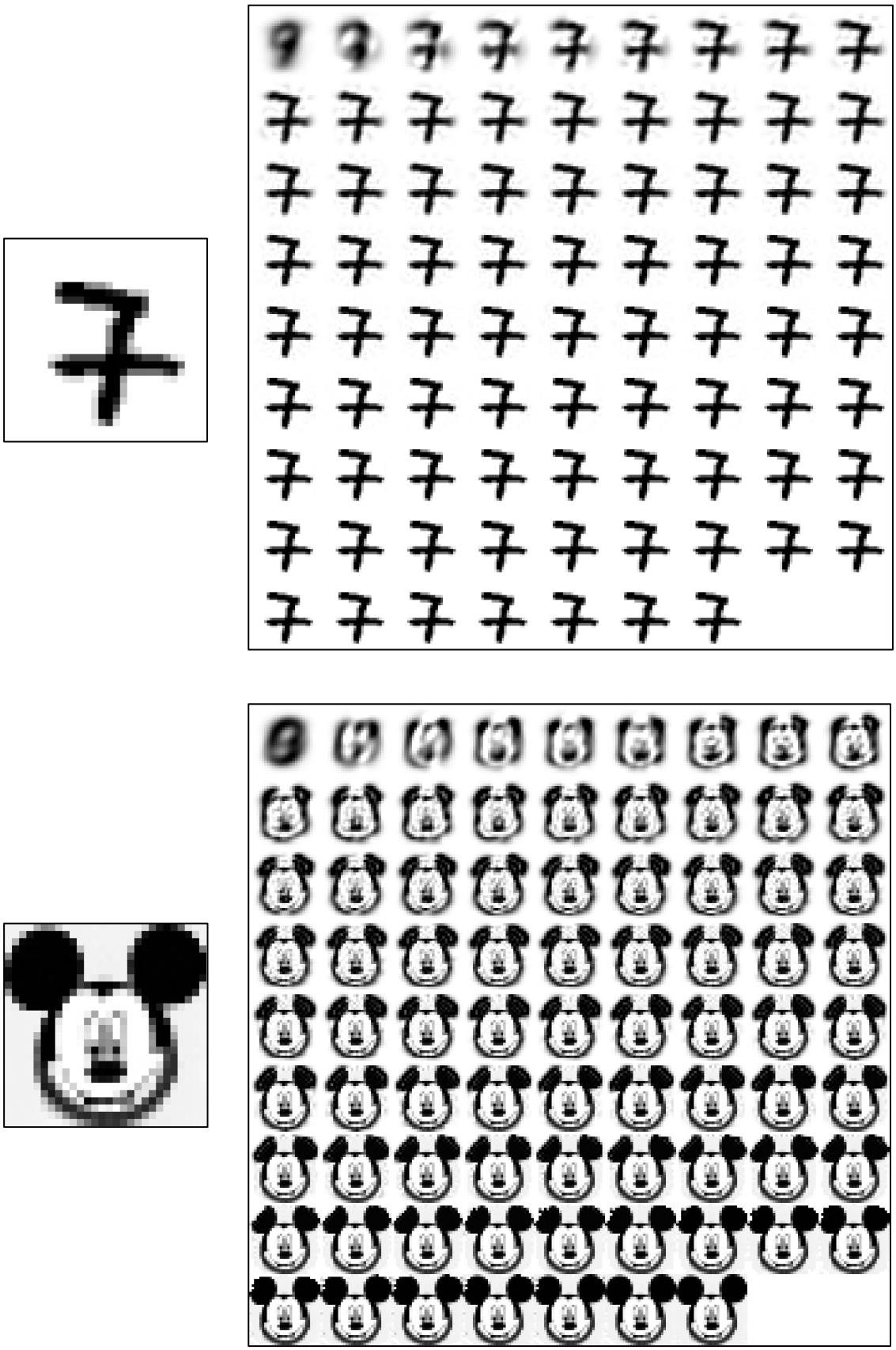

Figure 6. PCA reconstruction examples. The number of eigenvectors used for reconstructions are 2, 12 ,22, 32, …, 784 (ordered along the rows).

Evaluation Criteria

		Points
Correctness of the solution		50
Report (Contents, completeness, format, etc.)		40
Compliance to Submission Rules (Directory structure, file formats/naming, organization, etc.)		10
	TOTAL	**100**

Submission Guide

Submission Files
Submit a single compressed (.zip) file, named as name_surname.zip, to the Moodle. It should contain all source code files (under the \code directory), report (in PDF format, under the \report directory) and all other files if needed (under \misc directory)

File Naming
Name your report as name_surname.pdf. Name the main code which is used to run your assignment as assignmentX.py, where X is the assignment number and .py is the extension for Python, given as an example.

Late Submission Policy
Maximum delay is two days. Late submission will be graded on a scale of 50% of the original grade.

Mandatory Submission
Submission of assignments is mandatory. If you do not submit an assignment, you will fail the course.

Plagiarism
Leads to grade F and YÖK regulations will be applied