PROJECT PLAN

1. Student names. (The project is to be done in groups of 3 students.)

Ming Zhu, Xuda Lin, Jiawei Lyu.

2. [Up to 3 lines] Definition of the problem, possibly relevant to your interests.

We plan to investigate *Automatic Face Verification*, of which the task is to determine whether two images depict the same person or not.

3. [Up to 3 lines] Description of the dataset (or datasets) to be used. Datasets should be already publicly available, since there is not enough time for you to collect data. For possible datasets, see the course webpage.

We plan to use the dataset *Labeled Faces in the Wild.* The data set contains more than 13,000 images of faces detected by the Viola-Jones face detector. Each face has been labeled with the name of the person pictured.

4. URL where the above dataset(s) is(are) available.

http://vis-www.cs.umass.edu/lfw/

5. [Up to 5 lines] Which 3 machine learning algorithms are going to be used? (You should list 3 algorithms, e.g., SVM, Prank, Adaboost, etc.) **You are allowed to either implement this from scratch or use third-party code, e.g., liblinear for SVM.**

1, Principal Component Analysis (PCA);
2, Linear Discriminant Analysis (LDA);
3, Multiple Kernel Learning (MKL).

6. [Up to 5 lines] Cross-validation technique (e.g., training/validation/testing, k-fold cross-validation, bootstrapping). **You MUST implement this from scratch.**

10-fold cross-validation.

7. [Up to 10 lines] Which hyperparameter(s) is(are) going to be tuned. **You MUST implement this from scratch.**

1, top k eigenvalues of data matrix;
2, the number of Gaussian models K and the number of local image descriptors T;
3, kernel number p.

8. [Up to 15 lines] Description of the experimental results, e.g., plots of number of samples versus accuracy (you can use different subsets of the same dataset), regularization parameter versus accuracy, ROC curves, plots of different datasets, etc. **You MUST implement this from scratch.**

After 10 fold cross validation, two indicators will be employed to evaluate the three algorithms. The first one is the Receiving Operating Characteristic Equal Error Rate (ROC-EER)), which is the accuracy at the ROC operating point where the false positive and false negative rates are equal. The second one is the classification accuracy.

We will plot the curves of ROC-EER or accuracy versus different hyperparameters mentioned in the previous section and the ROC curves with the optimal hyperparameters.

9. Which programming language are you going to use? (Only MATLAB, C++, Java and Python are allowed.)

Python.

Advice: Do not spend too much time on things such as "understanding the data", "memory problems because your data is too big", etc. Only if you are already familiar with computer vision, brain data, natural language processing, big data, parallelism, etc. then you can make use of those things, but this will not imply that you will get a higher grade just based on that fact. In general, I would recommend to use easy-to-understand datasets, and smaller subsets of the data, for instance.