Exploration on Facial Expression Classification

Linjiajie Fang *ID:20382284, Ziliang Xiao ID:20378386, Wenyong Zhang ID:20377289, Zexiao Xue ID:20403674*

Dept. of Mathematics, HKUST

May 12, 2017

1 Inroduction

Compared to the extensively studied and applied Face Recognition (FR), Facial Expression Recognition (FER) still remains to be challenging mainly due to the variety and subtlety of expressions. Any slight changes of facial organs like eyes can generate quite different emotions. Even people themselves cannot identify the correct (target) emotion just from pictures, not mention to computers. Judith *et al.* (2004) compares the accuracy in judging different emotional expressions between the female and the male, finding that women rate is more correct target emotions than men in general, and the accuracy ranges from 50% to 84%. Considering this, we neglect the gender difference lying in the accuracy of expression recognition in our tests. Unlike face recognition in which most information of the image make a difference, only limited information of an image plays an important role in the expression recognition, others being redundant. Therefore, we eliminate the useless information such as hairs on the face before the test by cutting the original square image into oval image solely containing the face[1] .

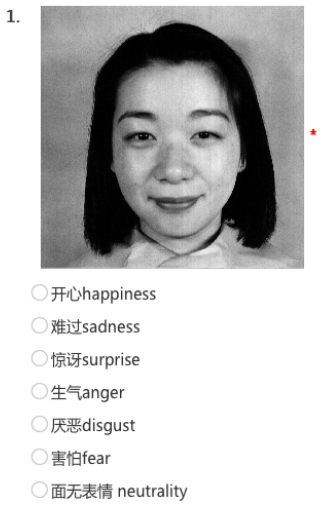

Figure 1: Questionnaire. 20 images are randomly selected for participators to match emotions. Finally 142 Questionnaires are collected

Naturally we think of shrinking the feature space to keep the most useful information. Firstly, we develop our linear regression approach, by using PCA to do dimensional reduction, and using the R^2 as classification criteria, the results turned out to be not accurate. Secondly, we use *Linear Discriminant Analysis* LDA and PDA to test the accuracy of identifying a certain emotion, but the result is also not satisfactory. Presumably aimlessly reducing the feature space does not make sense. Since the eyes and mouth usually dominate the changes of facial expression, we cut up each image from the middle to obtain two non-overlapping subblocks containing the eyes and mouth separately. Then we try different learning methods: *Logistic Regression* (Logit), *support vector machine* (SVM) and other methods to compare and improve the accuracy of emotion recognition as well as possible.

Meanwhile, we conduct a survey to capture the accuracy of recognizing female's facial expressions by people, mainly by HKUST students. We randomly pick twenty pictures from our database and ask the questionnaire taker to choose which expression does the

picture show among seven alternatives, namely anger, disgusting, fear, happy, neutral, sad, and surprise. We receive 142 pieces of feedback totally, and based on the feedback we calculate the recognition accuracy. The recognition accuracy is about 61%, for different expressions the recognition accuracy varies from 25% to 91%. The questionnaire can be viewed through following link: https://sojump.com/jq/13899647.aspx. Lastly, we compare the accuracy achieved by the computer with that from people, find our machine learning method outperforms the subjective cognition, thus is sensible to some extent.

2 Data Base

Our data base comes from 'The Japanese Female Facial Expression (JAFFE) Database' (http://www.kasrl.org/jaffe.html). We cut the original square image into oval image solely containing the face to remove redundant parts. The data base includes 213 emotion labeled images by ten people, $(x_i, y_i) \in G_i$, $|G_i| \approx 30$, $\{i = 1, 2, ..., 7\}$, where x_i is the matrix of the image and y_i is the emotion category. Each person provides 1 to 3 images for each emotion from 7 emotion categories: anger, disgusting, fear, happy, neutral, sad, and surprise.

Figure 2: Database

In our emotion recognition problem, we split the data into two parts: 140 images for training , 73 images for testing. Training data are 7 groups of emotion labled images $(x_i, y_i) \in G_i^{train} \subset G_i$, $\{i = 1, 2, ..., 7\}$. Each group contains 20 images, where 2 for each person, that is $|G_i^{train}| = 20$. Then we treat the rest as testing data. Our split is fair since both training data and test data include images from all emotions of each person in equal amount. This is to avoid over-learning on a certain emotion as well as insufficient-learning on another emotion.

[1]The transformation formula is $\frac{(y-128)^2}{80^2} + \frac{(x-128)^2}{60^2} = 1$.

3 Methodology

Our classification methods including:

- Linear Regression Approach
- LDA and Penalized LDA
- SVC with Gaussian Kernel
- Logistic Regression
- Bagging
- Random Forest
- Stochastic Gradient Descent SVC/Logistic

3.1 Our original linear regression approach

We consider implementing linear separation of data, and target to reduce the dimensions of feature space. Two methods are adopted and compared, one is *Principal Components Analysis* (PCA), the other one is *Linear Discriminant Analysis* (LDA).

In this scenario, we treat our samples in a different way from doing other methods. Rather than treat a image as a sample point, we treat a pixel position in the image as a sample point. The dimension of the sample points equals to the number of training picture, which is 20. The sample size $N = 160 \times 120$ since we have this many pixels for each image. Using this angle of sight, we do linear regression classifier and LDA classifier.

Our linear regression approach is built by the intuition that once a input image x is delivering "happy" for instance. The input image can be roughly represented by linear combination of training data which is labeled "happy", that is:

$$\text{input image } x \approx \beta_0 + \sum_{j=1}^{20} \beta_j x_j, x_j \in G_{happy}$$

In a view of pixels as sample points showed above, it becomes a regression problem:

$$\min_{\beta} ||p_i - \sum_{j=1}^{20} \beta_j x_{i,j}||^2 \text{ for each pixel } p_i, i \in \{1, 2, ..., N = 160 \times 120\}$$

No matter how β is fitted, the regression gives a criterion R^2 to indicate how well the regression fits. If we do linear regression of the input image x to all different groups of pictures, the R^2 of the regression should reach the maximum if the independent variables of the regression come from the same category "happy" as x does. Suppose $R^2(x, G_i)$ is the R^2 obtained by doing linear regression of x to the images in the group G_i. Our linear regression approach classifier is:

$$\hat{y} = \arg\max_k R^2(x, G_k), k \in \{1, 2, ..., 7\}$$

Here the number of independent variables in regression is 20, and we do not expect too many $\{x_j\}$, which may smooth the features of the emotions and give insignificant different R^2. So we do PCA on total number of $N = 160 \times 120$ samples, it turns out that the first 5 PCs of each emotion is enough to explain around 75% to 80% variance information for the corresponding group. Rather than doing regression for all 20 members in each group, we do regression for only the $PC_i^k, i = \{1, 2, 3, 4, 5\}$ in the k-th group. So as to pick out the PC that dose not explain the variation of input x, we choose lasso penalty in regression to magnify difference of R^2 between true and false group. So our classifier is:

$$\hat{y} = \arg\max_k R^2_{lasso}(x, \{PC_i^k\}_{i=1,...,5}), k \in \{1, 2, ..., 7\}$$

3.1.1 Algorithm

Step 1 Basic steps of PCA algorithm [14]

- Do PCA for each group. Choose the first five principal components $\{PC_i^k\}_{i=1,...,5} := \widetilde{G}_k$ as "template" for the k-th group, $k = \{1, 2, ..., 7\}$
- Note that PC_i are images visualized similar to Negative film.

Step 2 Given an un-labled input image x

- Treat each pixels $p_i \in x$ as dependent variable, and do regression with Lasso penalty for group $k = \{1, 2, ..., 7\}$:

  $$\min_\beta ||p_i - \sum_{j=1}^{5} \beta_j PC_{i,j}||^2 + \lambda||\beta||_1 \text{ for each pixel } p_i, i \in \{1, 2, ..., N\}$$

  and compute the R^2 denoted as $R^2_{lasso}(x, \widetilde{G}_k)$

- The predicted group is given by:

  $$\hat{y} = \arg\max_k R^2_{lasso}(x, \widetilde{G}_k), k \in \{1, 2, ..., 7\}$$

3.2 LDA and penalized LDA

Different from PCA simply maximizes between-class data separation, LDA tries to maximize between class variance and minimize within class variance at the same time. LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all K classes. Moreover, *Penalized Discriminant Analysis* is also tried to make contrast.

For Linear Discriminant Analysis, we use projected data. Based on the paper *(YAMBOR, WENDY S.: Analysis of PCA-Based and Fisher Discriminant-Based Image Recognition Algorithms)*. We project images from P dimensional space to K-1 dimensional space and here are steps of this method:1.create data matrix: combine all images as a N*P matrix (P is the number of pixels of each centered data).2. compute covariance matrix and get the first K-1 eigenvalues and orthogonal eigenvectors and order them $V = [v(1), v(2), \ldots, v(k-1)]$. Hence this is the projection matrix. 3.Project images into eigenspace, thus we get $\tilde{x} = V'x$.

After projection, we use N*(K-1)data to do the LDA and Penalized LDA. In LDA method, We just estimate the eigenvector which maximize the variance.
(LDA-L1): When the tuning parameter λ_k is large, some elements of the solution will be exactly equal to 0. σ_j is the within-class standard deviation for feature j, so features that vary more within each class undergo more penalization.

$$\max_{\beta_k}(\beta_k^T \hat{\Sigma}_b^k \beta_k - \lambda_k \sum_{j=1}^{p} |\hat{\sigma}_j \beta_{k,j}|) \text{ s.t. } \beta_k^T \tilde{\Sigma}_w \beta_k \leq 1$$

(LDA-FL): Besides properties of LDA-L1, this classifier is appropriate if the features are ordered on a line, and one believes that the true underlying signal is sparse and piecewise constant (a further penalization).

$$\max_{\beta_k}(\beta_k^T \hat{\Sigma}_b^k \beta_k - \lambda_k \sum_{j=1}^{p} |\hat{\sigma}_j \beta_{k,j}| - \gamma_k \sum_{j=2}^{p} |\hat{\sigma}_j \beta_{k,j} - \hat{\sigma}_{j-1} \beta_{k,j-1}|) \text{ s.t. } \beta_k^T \tilde{\Sigma}_w \beta_k \leq 1$$

Based on 73 samples as test data, we have 20-fold cross-validation. Figure 7 and **??** show the result from LDA, Figure 8 and **??** correspond to that of penalized LDA respectively.

3.3 SVC by Gaussian Kernel

We also use SVC with Gaussian Kernel,

$$k(x_i, x_j) = exp(-\gamma \parallel x_i - x_j \parallel^2)$$

The reason why we use Gaussian kernel is that our pixels in the image may diffuse randomly in different pictures. We assume the "featured" pixel that contains same important information would occur follow the normal distribution in different images. Suppose $C > 0$ is the penalty parameter of the error term used in determination of a separating hyperplane with the maximal margin in higher dimensional space by SVM. γ is the tuning parameter in Gaussian kernel. We use cross-validation to estimate optimal tuning parameters (C, γ).

4 Results

4.1 Subtracted PC_i and use R^2 as classification criterion

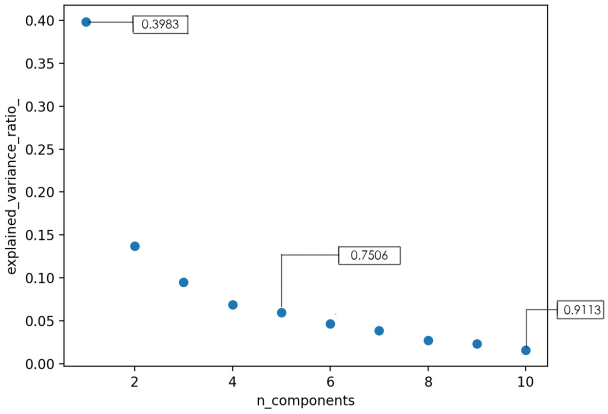

Figure 3: Principal components chosen for emotion "surprise". The first 5 principal components already explain over 75% variance of the training data. For other emotions we also choose the first 5 principal components as "template" to compare input image to a certain emotion.

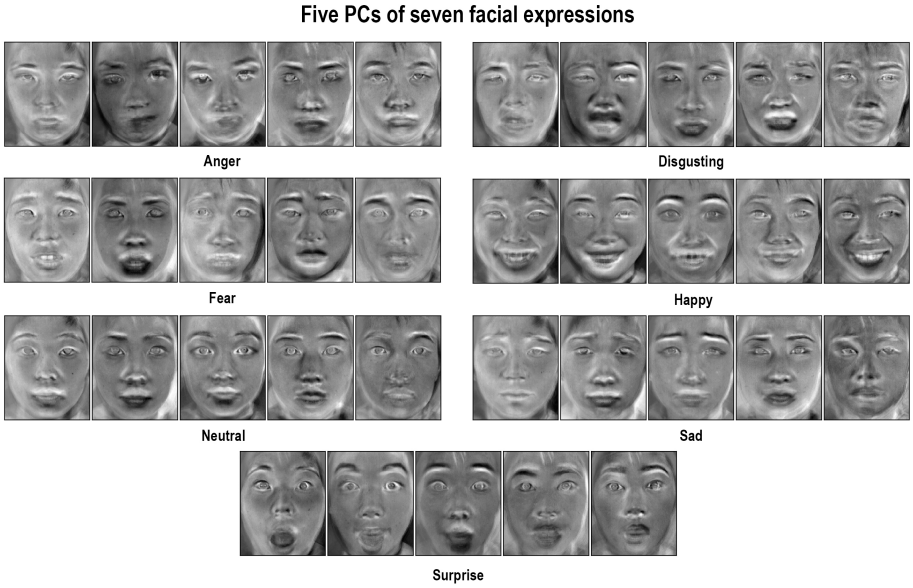

Figure 4: The first 5 principal components of each emotion. For each emotion, the $PC_1, PC_2, ..., PC_5$ are visualized from left to right. Principal components are similar to Negative film.

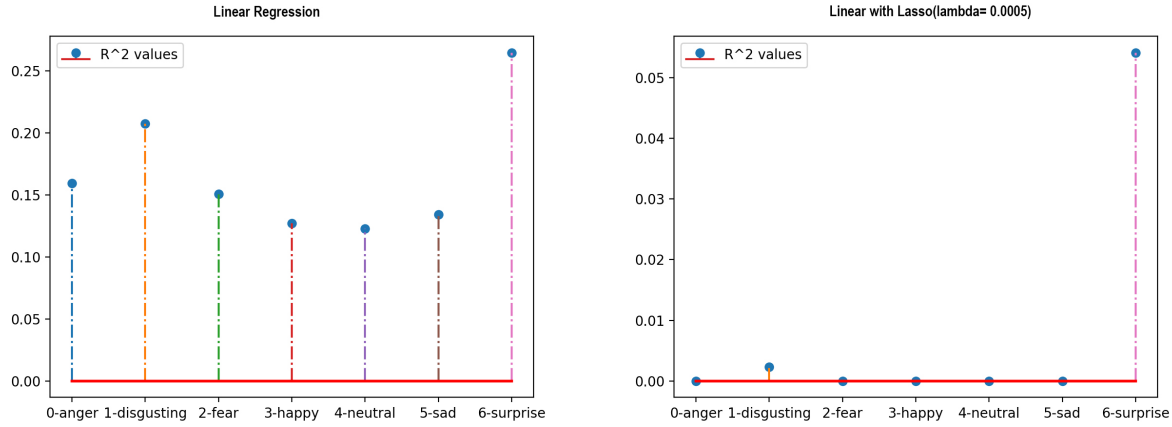

Figure 5: Classification criterion. The figure showed above is a example for classifying a input image. Linear regression gives that R^2 reach the maximum at emotion "surprise". After using Lasso penalty, the R^2 for wrong group decayed significantly and enlarge the difference between right and wrong groups. Since the largest R^2 is the criterion for classification. Here we classify the input to group "surprise".

4.2 LDA and PDA

In Figure 6, we draw pairwise plot of test data of LDA, we can easily see position of test data points in different dimensions. The confusion matrix and test accuracy are shown in Figure 7. In Penalized LDA, we choose the LDA-FL method and do a further penalization of features, through Figure 8, we can get the best K is 3 and we reduce the dimension from 6 to 3 and get the test accuracy as 0.45. It's clear that the accuracy has been enhanced through the process of penalization. However, one weakness of this way is that we throw too many other features away while doing the eigen-projection. This can be improved if we firstly make each image to be sparse.

Since the accuracy can be tolerated, this method is not bad for classifying with dimension reduced data.

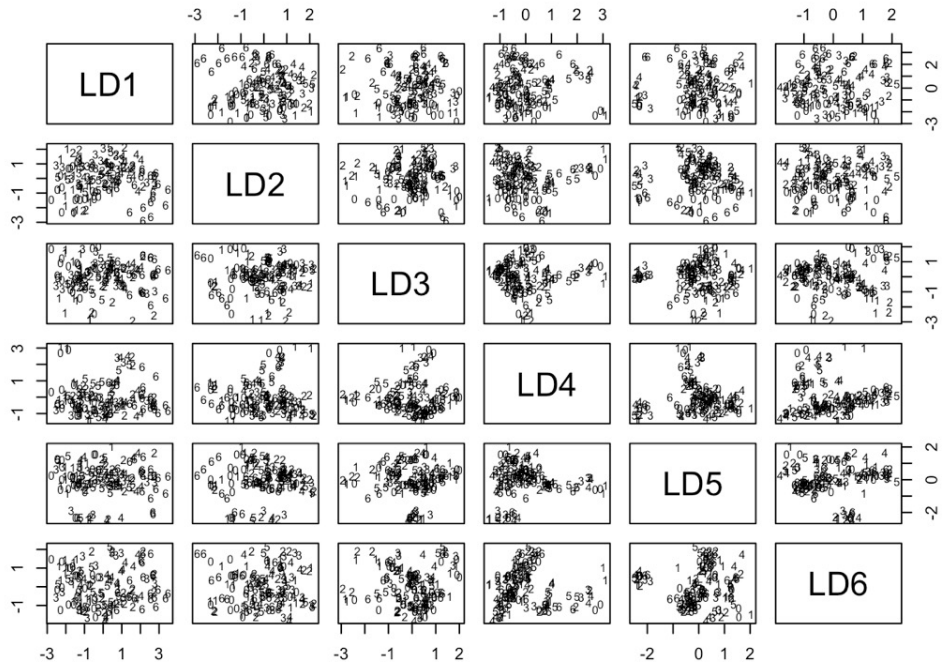

Figure 6: *We draw a plot of test data of LDA, the 6-D data is projected on 2-D subspace so we can see the position of test data points in different dimensions . The confusion matrix and test accuracy is shown in Figure 2 and corresponding describtions are listed under each pictures. In Penalized LDA, we choose the LDA-FL method to do a further penalization of features, through Figure 3, we can find out that the best K is 3 and we reduce the dimension from 6 to 3 and get the test accuracy as much as 0.45. It's clear that the accuracy has been enhanced through the process of penalization. However, one weakness of this way is that we throw too many other features away while doing the eigen-projection. This can be improved if we firstly make each image to be sparse. Since the accuracy can be tolerated, this method is not bad for classifying with dimension reduced data.*

Test Results of LDA and PDA

LDA	Anger	Disgusting	Fear	Happy	Neutral	Sad	Surprise
Anger	7	3	1	1	0	1	0
Disgusting	2	4	0	1	1	4	0
Fear	0	0	2	0	2	1	1
Happy	1	2	1	6	1	4	0
Neutral	0	0	2	1	4	1	0
Sad	0	0	1	0	2	0	1
Surprise	0	0	5	2	0	0	8
Correct Rate	0.4246575						

PDA	Anger	Disgusting	Fear	Happy	Neutral	Sad	Surprise
Anger	7	3	1	1	0	3	1
Disgusting	1	4	0	1	0	2	0
Fear	0	0	3	0	3	1	1
Happy	1	2	1	6	0	3	0
Neutral	0	0	1	1	5	2	0
Sad	1	0	1	0	2	0	0
Surprise	0	0	5	2	0	0	8
Correct Rate	0.4520548						

Figure 7: *The correct rate of LDA and PDA classifier for test data is about 42% and 45% respectively. In the confusion matrix, the true group index are shown in the column, and the number of images classified to a certain group indexed in the raw above. The numbers in the diagonal represent correct classification. Wrong classification is colored in red. After using PDA, the results updated only by two more images. Here: 0-anger, 1-disgusting, 2-fear, 3-happy, 4-neutral, 5-sad, 6-surprise*

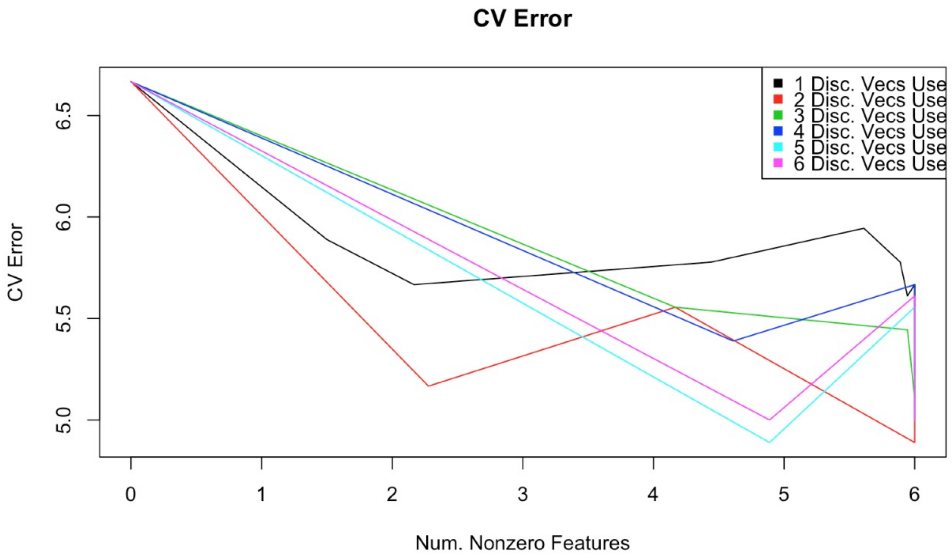

Figure 8: *The figure shows that the best choice of K is 3. The cross-validation error with respect to the first and the second eigenvector both get minimum value when K is slightly bigger than 2 but smaller than 3, so we take the smallest integer which is bigger than the minimum point. (If the CV error w.r.t first two eigenvectors can not get to minimum value from 0 to 6, we will then take vectors with smaller eigenvalue into consideration.)*

4.3 Choosing Tuning Parameters in updated SVC

Cross validation is very essential step to select the tuning hyper-parameters (C and γ). We use following grid search method: C ranges from 10^{-2} to 10^{10}, γ from 10^{-10} to 10^{2}, and 13×13=169 points for iterations.

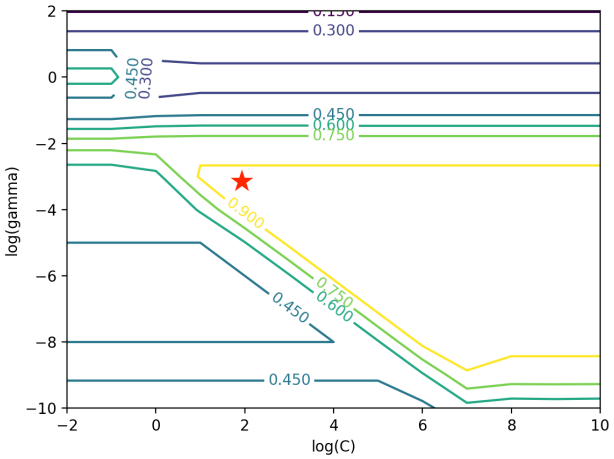

Figure 9: Contour map of log(C) and log(γ). Each line represents parameter pair $(log(C), log(\gamma))$ sharing same accuracy from 20-fold cross-validation on training data. The parameters we finally chosen are $(log(C), log(\gamma)) = (2, -3)$ (The red star), in the region where accuracy over 90% is reached.

4.4 Number of Folds Chosen in Cross-Validation

In order to let the model be more convincing and make a trade-off between accuracy and train time, we compare different number of folds with training accuracy in the cross-validation:

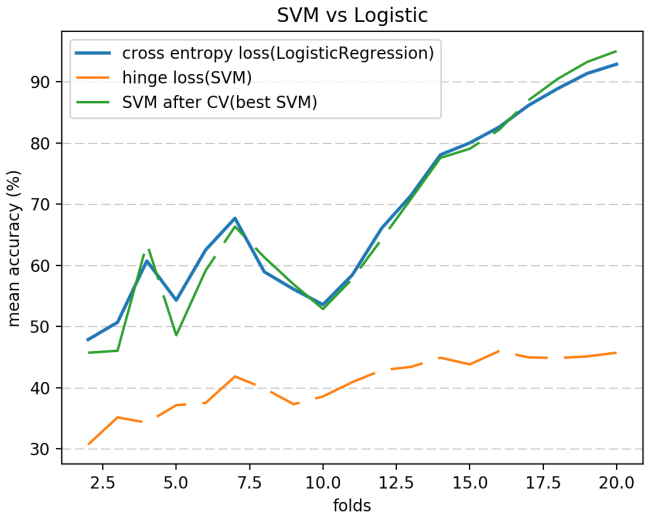

Figure 10: Three models comparison whole face

We are able to select the updated SVM from this figure.

With the increasing of the k-folds, the accuracy of classification also goes up. Although our updated SVM performs quite well, we also saw the huge power of logistic regression which is relied on the cross-entropy loss.

4.5 Test Accuracy

The accuracy of all the method we use is show below. The highest accuracy among these methods is the updated SVM, which achieves 0.893. The linear regression, LDA, PDA, and classic SVM are not perform well, the accuracy of them are less than the survey results. The accuracy of tree methods which contain bagging and random forest, and the SGD method which including SGDsvm and SGDlog is around 0.6.

Correct Rate

Methods	20-fold CV	Test Data
Linear Regression	-	0.461
LDA	0.434	0.428
PDA	0.486	0.453
Classic SVM	0.428	0.521
Logistic	0.928	0.890
Updated SVM	0.951	0.893
Bagging	0.536	0.629
Random Forest	0.586	0.653
SGDsvm	0.536	0.597
SGDlog	0.579	0.610
Survey(142 copies)	-	0.614

Random State	Bagging	Ran.Forest	SGDsvm	SGDlog
Rates	0.712	0.740	0.575	0.616
	0.671	0.562	0.575	0.589
	0.589	0.685	0.616	0.589
	0.630	0.699	0.630	0.644
	0.548	0.658	0.630	0.658
	0.630	0.616	0.644	0.617
	0.616	0.671	0.589	0.616
	0.616	0.616	0.671	0.685
	0.616	0.630	0.479	0.534
	0.658	0.658	0.562	0.548
Mean Rate:	0.629	0.653	0.597	0.610

5 Comparing Emotion Information Contained in Eyes and Mouth

Another idea comes that the accuracy comparison between eyes and mouth. We split train and test data into two parts, with pixels 90*120 and 70*120.

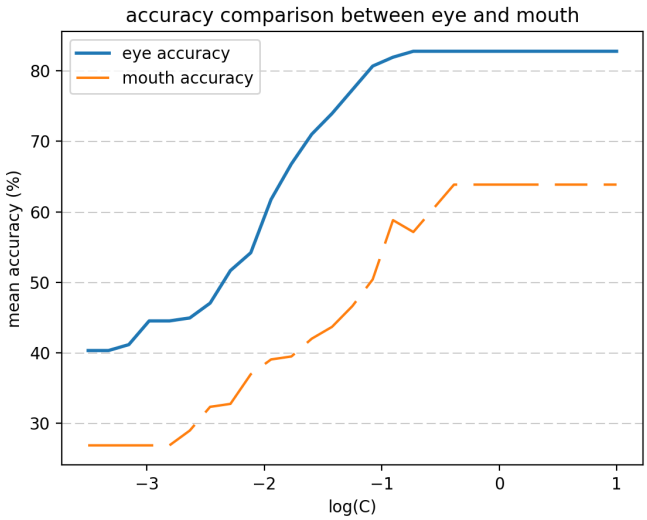

Figure 11: Accuracy comparison between eyes and month

The result of image shows that eyes outperform in accuracy than mouth under same margin constrain C. This phenomenon illustrates that eyes could explain more information in facial express recognition(actually much better than mouth).

6 Prediction——realistic application

In order to test our classifiers, we considered doing prediction using pictures sampling from one of our group members——Zhang, Wenyong(ZWY). And we choose the best two classifiers: updated SVC with Gaussian kernel and logistic classifier. After eliminating the background noise, we are able to get the following samples:

Figure 12: Seven different emotions of ZWY. From left to right are emotions: anger, disgusting, fear, happy, neutral, sad, surprise, respectively.

Each image represents one emotion. Applying our trained SVC classifier(updated SVM) and trained logistic classifier, the prediction gives :

Emotion	anger	disgusting	fear	happy	neutral	sad	surprise
SVC(G.K.)	fear	disgusting	surprise	disgusting	anger	fear	surprise
Logistic	fear	disgusting	surprise	disgusting	anger	fear	surprise

Unfortunately, our classifiers only do two correct classifications.

Another trying is to eliminate the mouth part of ZWY with updated mouth-deleting trained data:

Figure 13: Cutting images with high contrast effect. Emotions(left to right: anger, disgusting, fear, happy, neutral, sad, surprise, respectively.

Applying our trained classifiers again, a little bit better result is obtained than the previous prediction.

Emotion	anger	disgusting	fear	happy	neutral	sad	surprise
SVC(G.K.)	anger	disgusting	disgusting	disgusting	disgusting	disgusting	surprise
Logistic	anger	disgusting	disgusting	disgusting	disgusting	disgusting	surprise

We got 3 correct predictions from 7 emotions. Other emotions result in "disgusting" group. The improvement in the result indicates that the mouth part of our volunteer ZWY may confuses the classifier.

7 Conclusions

- Generally speaking, we find that SVM after cross-validation and logistic model with cross-entropy get best accuracy which is as much as 0.89 on test data. It seems erratic that machine accuracy is much higher than human accuracy in the questionnaire(0.614). This phenomenon is owe to the fact that our train and test data are from 10 women, and we just predict facial expressions among these 10 people. Our classifier sufficiently learns their features and even includes some features that are private rather than general. This might explain the result why our classifier are not accurately working on our group member ZWY. Because some features of appearance of ZWY is very different from women in the training data, which disturbs the emotion recognition. Since our training sample only contains 140 samples from 10 people, a larger data set with more people participating may train a better classifier.
- The stochastic gradient descend algorithm which is applied to SVC and logistic way does not behave better than tree methods which includes bagging and random forest.

After separating images into eyes and mouths, we find that test accuracy with eyes is higher than that of mouth, which does also confirm the old saying: *Eyes can reflect ones heart.*

References

[1]Judith A.Hall and David Matsumoto. 2004. Gender Differences in Judgements of Multiple Emotions From Facial Expression. *British Journal of Cancer*, **8**, 112.

[2]Wendy S. Yambor. 2000. Analysis of PCA-Based and Fisher Discriminant-Based Image Recognition Algorithms.

[14]YAMBOR, WENDY S. : Analysis of PCA-Based and Fisher Discriminant-Based Image Recognition Algorithms, http://www.cs.colostate.edu/evalfacerec/papers/tr 00-103.pdf.

[8]TURK, M.—PENTLAND, A. : Eigenfaces for Recognition, Journal of Cognitive Neuroscience 3 No. 1 (1991), 71-86.

[19]HSU, C. W.—CHANG, C. C.—LIN, C. J. : A Practical Guide to Support Vector Classication, 2008 http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf.