Pattern Recognition
CS669

Course Instructor : Dr. Dileep A. D.

ASSIGNMENT 4: PART 1

Gaussian Mixture Models on
Reduced Dimensional Representations
using
Principal Component Analysis

Group Number 8

Aman Khandelwal	B16007
Amrendra Singh	B16010
Bharat Lodhi	B16015

Contents

List of Tables

1. Problem Description

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs on the basis of a training set of data containing observations (or instances) whose category membership is known.

Data-sets:

- 3 class scene image dataset

Classifiers to be built:

- Build Bayes classifier using Gaussian mixture model (GMM) with 1, 2, 4 and 8 mixtures on the reduced dimensional representations of data-set2 obtained using PCA.

2. Solution Approach

1 Principal Componenent Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities, each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

Procedure

1. Take all the training examples of all the classes.
2. Compute the co-variance matrix.
3. Perform eigen analysis to obtain d eigenvalues and corresponding eigenvectors where d is the dimension of one feature vector.
4. Choose L eigen vectors corresponding to the L leading eigen values where L is the number of dimensions in which we our data is reduced to from d dimensions.
5. Project each of the data points (training as well as test) onto these L eigen vectors to obtain the L principal components.

After reducing the dimension of the 32-dimensional BoVW representation of our images, GMM based classification was carried out on the reduced dimensional dataset.

We have carried out the GMM based classification for $1, 2, 4, 8$ mixture components.

3. Results

1 Classification using GMM on Color Histograms Results

	C = 1	C = 2	C = 4	C = 8	C = 16	C = 32
Accuracy	46.0%	24.66%	40.0%	45.33%	40.66%	47.99%
mean precision	0.5146	0.2368	0.3852	0.4769	0.4103	0.4908
mean recall	0.46	0.2466	0.4155	0.4533	0.4066	0.4799
mean F-Measure	0.4419	0.2237	0.3878	0.4397	0.4061	0.4779

Table 3..1. GMM Results for Color Histograms

2 Classification using GMM on BOVW representaion

	Bayou	Chalet	Creek
Bayou	9	9	32
Chalet	7	7	36
Creek	12	2	36

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3214	0.3889	0.3462
Recall	0.18	0.14	0.72
F-Measure	0.2308	0.2059	0.4675

(b) Analysis

Table 3..2. Confusion Matrix and Analysis: C = 1

	Bayou	Chalet	Creek
Bayou	26	11	13
Chalet	8	15	27
Creek	18	4	28

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.5	0.5	0.4118
Recall	0.52	0.3	0.56
F-Measure	0.5098	0.375	0.4746

(b) Analysis

Table 3..3. Confusion Matrix and Analysis: C = 2

	Bayou	Chalet	Creek
Bayou	23	20	7
Chalet	10	34	6
Creek	15	24	11

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4792	0.4359	0.4583
Recall	0.46	0.68	0.22
F-Measure	0.4694	0.5312	0.2973

(b) Analysis

Table 3..4. Confusion Matrix and Analysis: C = 4

	Bayou	Chalet	Creek
Bayou	25	13	12
Chalet	16	29	5
Creek	10	12	28

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4902	0.537	0.6222
Recall	0.5	0.58	0.56
F-Measure	0.495	0.5577	0.5895

(b) Analysis

Table 3..5. Confusion Matrix and Analysis: C = 8

	Bayou	Chalet	Creek
Bayou	31	15	4
Chalet	21	22	7
Creek	15	18	17

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4627	0.4	0.6071
Recall	0.62	0.44	0.34
F-Measure	0.5299	0.419	0.4359

(b) Analysis

Table 3..6. Confusion Matrix and Analysis: C = 16

	Bayou	Chalet	Creek
Bayou	32	12	6
Chalet	24	15	11
Creek	24	16	10

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4	0.3488	0.3704
Recall	0.64	0.3	0.2
F-Measure	0.4923	0.3226	0.2597

(b) Analysis

Table 3..7. Confusion Matrix and Analysis: C = 32

	C = 1	C = 2	C = 4	C = 8	C = 16	C = 32
Accuracy	34.67	46.0	45.33	54.67	46.67	38.0
mean precision	0.3522	0.4706	0.4578	0.5498	0.4899	0.3731
mean recall	0.3467	0.46	0.4533	0.5467	0.4667	0.38
mean F-Measure	0.3014	0.4531	0.4326	0.5474	0.4616	0.3582

Table 3..8. GMM Results for BOVW

3 Principal Component Analysis

3.1 Variation of Eigen-Values

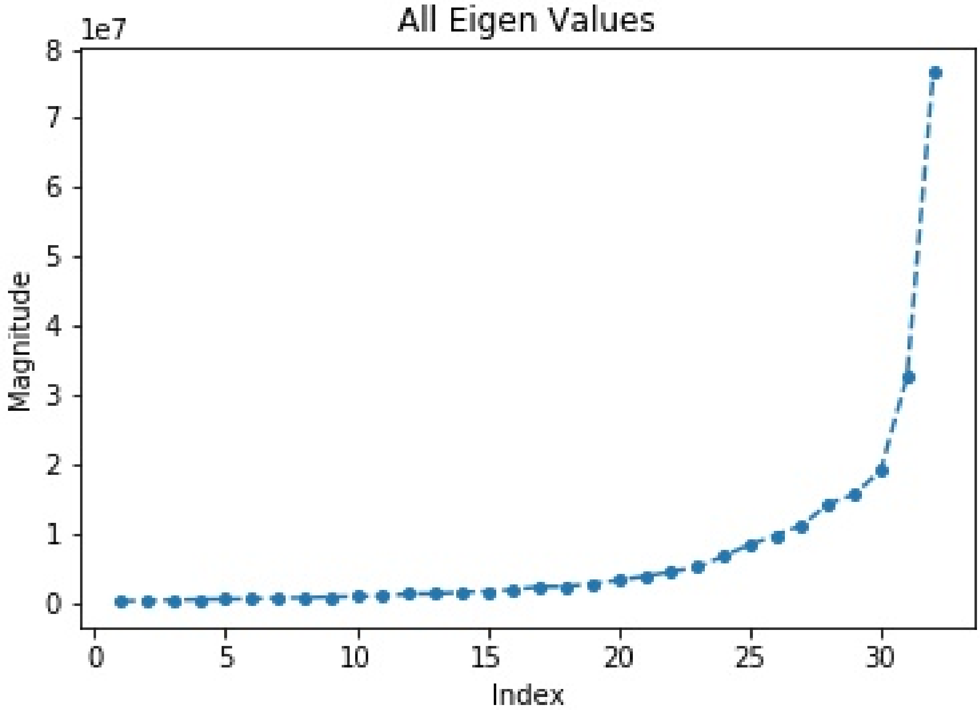

Figure 3..1. Variation of Eigen Values

We observe that around 5 - 10 eigenvalues represent almost 80 percent of variation in our data, while other eigenvalues don't contribute much. We take the most significant L eigen vectors corresponding to the most significant L eigen values.

The values of L considered are $2, 5, 10, 15, 21$.

3.2 Confusion Matrix, Precision, Recall and F-measure

L = 2

	Bayou	Chalet	Creek
Bayou	9	2	39
Chalet	6	4	40
Creek	5	6	39

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.45	0.3333	0.3305
Recall	0.18	0.08	0.78
F-Measure	0.2571	0.129	0.4643

(b) Analysis

Table 3..9. Confusion Matrix and Analysis: L = 2 , C = 1

	Bayou	Chalet	Creek
Bayou	27	14	9
Chalet	19	19	12
Creek	27	17	6

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3699	0.38	0.2222
Recall	0.54	0.38	0.12
F-Measure	0.439	0.38	0.1558

(b) Analysis

Table 3..10. Confusion Matrix and Analysis: L = 2 , C = 2

	Bayou	Chalet	Creek
Bayou	22	17	11
Chalet	18	17	15
Creek	21	17	12

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3607	0.3333	0.3158
Recall	0.44	0.34	0.24
F-Measure	0.3964	0.3366	0.2727

(b) Analysis

Table 3..11. Confusion Matrix and Analysis: L = 2 , C = 4

	Bayou	Chalet	Creek
Bayou	17	21	12
Chalet	13	24	13
Creek	14	21	15

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3864	0.3636	0.375
Recall	0.34	0.48	0.3
F-Measure	0.3617	0.4138	0.3333

(b) Analysis

Table 3..12. Confusion Matrix and Analysis: L = 2 , C = 8

	C = 1	C = 2	C = 4	C = 8
Accuracy	34.67%	34.67%	34.0%	37.33%
Mean Precision	0.3713	0.324	0.3366	0.375
Mean Recall	0.3467	0.3467	0.34	0.3733
Mean F-Measure	0.2835	0.325	0.3353	0.3696

Table 3..13. PCA Results : L = 2

L = 5

	Bayou	Chalet	Creek
Bayou	7	3	40
Chalet	4	5	41
Creek	12	6	32

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3043	0.3571	0.2832
Recall	0.14	0.1	0.64
F-Measure	0.1918	0.1562	0.3926

(b) Analysis

Table 3..14. Confusion Matrix and Analysis: L = 5 , C = 1

	Bayou	Chalet	Creek
Bayou	32	14	4
Chalet	16	22	12
Creek	15	24	11

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.5079	0.3667	0.4074
Recall	0.64	0.44	0.22
F-Measure	0.5664	0.4	0.2857

(b) Analysis

Table 3..15. Confusion Matrix and Analysis: L = 5 , C = 2

	Bayou	Chalet	Creek
Bayou	17	10	23
Chalet	13	9	28
Creek	9	13	28

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4359	0.2812	0.3544
Recall	0.34	0.18	0.56
F-Measure	0.382	0.2195	0.4341

(b) Analysis

Table 3..16. Confusion Matrix and Analysis: L = 5 , C = 4

	Bayou	Chalet	Creek
Bayou	18	18	14
Chalet	13	25	12
Creek	5	27	18

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.5	0.3571	0.4091
Recall	0.36	0.5	0.36
F-Measure	0.4186	0.4167	0.383

(b) Analysis

Table 3..17. Confusion Matrix and Analysis: L = 5 , C = 8

	C = 1	C = 2	C = 4	C = 8
Accuracy	29.33%	43.33%	36.0%	40.67%
Mean Precision	0.3149	0.4273	0.3572	0.4221
Mean Recall	0.2933	0.4333	0.36	0.4067
Mean F-Measure	0.2469	0.4174	0.3452	0.4061

Table 3..18. PCA Results : L = 5

L = 10

	Bayou	Chalet	Creek
Bayou	8	4	38
Chalet	4	7	39
Creek	14	5	31

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3077	0.4375	0.287
Recall	0.16	0.14	0.62
F-Measure	0.2105	0.2121	0.3924

(b) Analysis

Table 3..19. Confusion Matrix and Analysis: L = 10 ,
C = 1

	Bayou	Chalet	Creek
Bayou	27	12	11
Chalet	14	23	13
Creek	18	19	13

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4576	0.4259	0.3514
Recall	0.54	0.46	0.26
F-Measure	0.4954	0.4423	0.2989

(b) Analysis

Table 3..20. Confusion Matrix and Analysis: L = 10 , C = 2

	Bayou	Chalet	Creek
Bayou	20	12	18
Chalet	16	4	30
Creek	15	7	28

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3922	0.1739	0.3684
Recall	0.4	0.08	0.56
F-Measure	0.396	0.1096	0.4444

(b) Analysis

Table 3..21. Confusion Matrix and Analysis: L = 10 , C = 4

	Bayou	Chalet	Creek
Bayou	13	19	18
Chalet	7	30	13
Creek	16	18	16

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3611	0.4478	0.3404
Recall	0.26	0.6	0.32
F-Measure	0.3023	0.5128	0.3299

(b) Analysis

Table 3..22. Confusion Matrix and Analysis: L = 10 , C = 8

	C = 1	C = 2	C = 4	C = 8
Accuracy	30.67%	42.0%	34.67%	39.33%
Mean Precision	0.3441	0.4116	0.3115	0.3831
Mean Recall	0.3067	0.42	0.3467	0.3933
Mean F-Measure	0.2717	0.4122	0.3167	0.3817

Table 3..23. PCA Results : L = 10

L = 15

	Bayou	Chalet	Creek
Bayou	8	9	33
Chalet	5	12	33
Creek	13	6	31

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3077	0.4444	0.3196
Recall	0.16	0.24	0.62
F-Measure	0.2105	0.3117	0.4218

(b) Analysis

Table 3..24. Confusion Matrix and Analysis: L = 15 ,
C = 1

	Bayou	Chalet	Creek
Bayou	25	19	6
Chalet	14	33	3
Creek	15	26	9

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.463	0.4231	0.5
Recall	0.5	0.66	0.18
F-Measure	0.4808	0.5156	0.2647

(b) Analysis

Table 3..25. Confusion Matrix and Analysis: L = 15 ,
C = 2

	Bayou	Chalet	Creek
Bayou	16	23	11
Chalet	10	26	14
Creek	11	18	21

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4324	0.3881	0.4565
Recall	0.32	0.52	0.42
F-Measure	0.3678	0.4444	0.4375

(b) Analysis

Table 3..26. Confusion Matrix and Analysis: L = 15 ,
C = 4

	Bayou	Chalet	Creek
Bayou	19	15	16
Chalet	6	25	19
Creek	12	17	21

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.5135	0.4386	0.375
Recall	0.38	0.5	0.42
F-Measure	0.4368	0.4673	0.3962

(b) Analysis

Table 3..27. Confusion Matrix and Analysis: L = 15 ,
C = 8

	C = 1	C = 2	C = 4	C = 8
Accuracy	34.0%	44.67%	42.0%	43.33%
Mean Precision	0.3572	0.462	0.4257	0.4424
Mean Recall	0.34	0.4467	0.42	0.4333
Mean F-Measure	0.3147	0.4204	0.4166	0.4334

Table 3..28. PCA Results : L = 15

L = 21

	Bayou	Chalet	Creek
Bayou	9	5	36
Chalet	6	5	39
Creek	15	3	32

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.3	0.3846	0.2991
Recall	0.18	0.1	0.64
F-Measure	0.225	0.1587	0.4076

(b) Analysis

Table 3..29. Confusion Matrix and Analysis: L = 21 ,
C = 1

	Bayou	Chalet	Creek
Bayou	9	24	17
Chalet	3	16	31
Creek	10	16	24

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4091	0.2857	0.3333
Recall	0.18	0.32	0.48
F-Measure	0.25	0.3019	0.3934

(b) Analysis

Table 3..30. Confusion Matrix and Analysis: L = 21 ,
C = 2

	Bayou	Chalet	Creek
Bayou	20	22	8
Chalet	11	33	6
Creek	17	18	15

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4167	0.4521	0.5172
Recall	0.4	0.66	0.3
F-Measure	0.4082	0.5366	0.3797

(b) Analysis

Table 3..31. Confusion Matrix and Analysis: L = 21 ,
C = 4

	Bayou	Chalet	Creek
Bayou	24	13	13
Chalet	14	25	11
Creek	20	11	19

(a) Confusion Matrix

	Bayou	Chalet	Creek
Precision	0.4138	0.5102	0.4419
Recall	0.48	0.5	0.38
F-Measure	0.4444	0.5051	0.4086

(b) Analysis

Table 3..32. Confusion Matrix and Analysis: L = 21 ,
C = 8

	C = 1	C = 2	C = 4	C = 8
Accuracy	30.67%	32.67%	45.33%	45.33%
mean precision	0.3279	0.3427	0.462	0.4553
mean recall	0.3067	0.3267	0.4533	0.4533
mean F-Measure	0.2638	0.3151	0.4415	0.4527

Table 3..33. PCA Results : L = 21

3.3 Variation of Accuracy with L and number of GMM components

	C = 1	C = 2	C = 4	C = 8
L = 2	34.67%	34.67%	34.0%	37.33%
L = 5	29.33%	43.33%	36.0%	40.67%
L = 10	30.67%	42.0%	34.67%	39.33%
L = 15	34.0%	44.67%	42.0%	43.33%
L = 21	30.67%	32.67%	45.33%	45.33%

Table 3..34. Variation of Accuracy with L and C

We observe that in general that accuracy increases with increase in L and the number of GMM components although there are a few spikes in occasionally.

4. Conclusion & Inferences

- Firstly, on average, accuracy across all clusters for a given L increases with increase in L.
- The essence of PCA is that it doesn't discard the features but provides us with reduced number of features which are combination of all the original features in the way, capturing the variation present in data.
- We observe that we obtain a comparatively decent accuracy as compared to the original 32-dimensional BOVW representation.
- We observe that we obtain almost similar accuracies with PCA in comparison to the analysis using the BOVW or the Color Histogram representations. Even with less features we obtain similar results, although not very high accuracy is seen.
- The best results were obtained using the GMM based classification on the 32-dimensional BOVW representations.

Advantage of dimension reduction :

- The more dimensions we take means more relationships between given features are taken, this may overfit our model but taking the most relevant information from all features increases the chance of our model not getting overfit.
- All the features don't always contain much additional information about our system. Our system may be expressed sufficiently in lesser number of features which have significant information about our system. This fact is utilized and we represent our system quite efficiently in lesser dimensions.
- This saves us a lot of space as well as computational costs for us because of the reduced size of our data representation.

Bibliography

[1] Bayes Classifier
https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysi

[2] Bayes Classifier
https://onlinecourses.science.psu.edu/stat505/node/49/

[3] Bayes Classifier
https://medium.com/@aptrishu/understanding-principle-component-analysis-e32be0

[4] K-Means clustering
https://en.wikipedia.org/wiki/K-means_clustering
https://www.geeksforgeeks.org/k-means-clustering-introduction/
https://towardsdatascience.com/understanding-k-means-clustering-\
in-machine-learning-6a6e67336aa1

[5] Bayes Classifier
https://en.wikipedia.org/wiki/Bayes_classifier

[6] Statistical Classification
https://en.wikipedia.org/wiki/Statistical_classification

[7] Naive Bayes Classifier
https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[8] Matplotlib Contours
https://matplotlib.org/api/_as_gen/matplotlib.pyplot.contour.
html

[9] Stack Overflow
https://stackoverflow.com

[10] K-Nearest Neighbor Method https://stats.stackexchange.com/
questions/252852/k-value-vs-accuracy-in-knn

[11] Baum Welch Algorithm https://en.wikipedia.org/wiki/Baum\OT1\
textendashWelch_algorithm