Principal Component Analysis and K-Means Clustering Biometrics- Project 2

Madhusudan Govindraju
University of Florida
Email: madhusudangr@ufl.edu

***Abstract*—This papers gives a detailed report on how the problem statement has been understood and the steps and assumptions taken to solve the problem. In this paper, steps undertaken to implement Principal Component Analysis(PCA) and KMeans clustering for face recognition has been elaborated, along with their performance evaluations. The performance evaluations for PCA used in this paper are Receiver Operator Characteristics curve, Cumulative Match Characteristics Curve & Genuine/Imposter Probablity Distributions. These plots are compared for different number of coefficients used in the PCA algorithm. The KMeans clustering algorithm has been implemented for gender a soft biometric classification example. Various external and internal characteristic indices have been used to evaluate the gender based soft biometric classification.**

I. Introduction

The image which is used to test against the system's database of image is called the probe and the images which make up the database is called the gallery images. The complete project can be broadly divided in to two; one, PCA for face detection step ;two; kmeans for soft biometric classification.

Principal Component Analysis(PCA) was first used for face recognition in [2], most algorithms available perform matching in two steps ; step one, project the images into the subspace; step two, measure the similarity and classify in the subspace. In [2] they demonstrate "that the any particular face can be represented in a best coordinate system that they term eigenpicture". The method to calculate this eigen picture is described in detail in the subsection *"Principal Component Analysis: EigenFaces method"* under the Method of Implementation. The performance of the system is evaluated using the CMC curve, ROC curve, genuine and imposter distribution curves. The curves obtained for the system implemented are described in detail in later sections.

Clustering is an unsupervised method. In other words, the labels are not given to the clustering algorithm. The clustering algorithm tries to find a common point and classifies the data.The KMeans clustering is a flat clustering approach. It is initialized with a local centroid and iteratively tries to minimize the cost function. One of the important step in data clustering is feature selection. So using more information that describes the data perfectly helps in producing a better set of clusters. The cost function and the steps to iteratively minimize the objective function is described in detail in the subsection *"K Means Clustering"*, under the "Methods of Implementation" section. The results of clustering algorithm is verified for correctness using internal validity criteria and external validity criteria. In out implementation we check the validity using the Calinski-Harabasz Index, Davies-Bouldin Index and Silhouette Width Criterion Index as the internal validity criteria; F-measure and Entropy as external validity criteria.

II. The Data

Fig. 1. Sample images From the GallerySet

Fig. 2. Sample images From the ProbeSet

Fig. 3. Sample images From the ProbeSet

The input data is a set of 3 images from 100 individuals. A total of 300 images. Each input image is of size 50x50 and

is of uint8 type. A sample of the images from the gallery and probe set is shown in the figure 3. The data set is divided into Gallery set and Probe Set. The first image from each individual is given as the gallery image and it corresponds to the training data. The other two images from each individual is given as the probe images and it corresponds to the testing data. We train the PCA for Face recognition algorithm with these 100 gallery images. We test the algorithm with the 200 probe images. From the figure 3 we can understand that the input faces are preprocessed and ready to use. If the images are not preprocessed we need to center & align the faces, apply illumination normalization effects to the database before using it in out system.

Before using the data we have to reshape the image into a 2500x1vector. This is in accordance with the PCA algorithm mentioned in the paper [2]. For eg., the gallery images will be aligned into a 2500x100 vector , where each column is a different sample.

III. Method of Implementation

The steps to implement the PCA for face detection and the Kmeans for gender classification are explained in the following subsection.

A. Principal Component Analysis: EigenFaces method

The steps to calculate the Eigen faces and the corresponding coefficients is as follows [2],

(a) The face images are loaded from the data sets and rearranged to form a vector of size 2500xM where M is the number of images. Each image was originally 50x50 which has been rearranges to a one dimensional vector 2500x1, This is Γ

(b) Find the mean image Ψ

(c) Subtract the mean image from every image to get the $\Phi(i)$ and we group all the $\Phi(i....m)$ together to get the matrix A.

(d) Now if we find the covariance of A to find the eigenvectors and eigenvalues . The resulting matrix is very large, instead of finding the covariance matrix $A \times A^T$ we find the matrix $A^T \times A$ and then compute the eigenvectors(v_i)

$$A^T A v_i = \mu_i v_i$$

(e) We sort the eigenvalues(μ) in descending order and then pick the top 3(to top 100 in steps) eigenvectors corresponding to the top 3 (to top 100 in steps) eigenvalues, and calculate the eigenvectorsu_i of the AA^T matrix using the following formula,

$$u_i = A v_i$$

(f) This u_i vector is rearranged in a 50×50 matrix to get the eigenfaces.

(g) The weights are calculated by multiplying the top N required eigenfaces with the matrix A.

Using the top 3 eigen vectors we rearrange the vectors into images of size 50x50 and that is available in the figurre 4. The first principal component or the first eigen vector corresponds to illumination varianec. And thus to normalize the effect of illumination, experts mostly drop the first eigen vector.

Fig. 4. Top 3 Eigen faces

B. K Means Clustering

In this type of clustering we start with a random point or centroid in this case and iteratively loop to proceed in the direction to reduce the cost function. The disadvantage of this algorithm is that we may reach a local minimum and it may not be anywhere near the global minimum, and if we don't know the data's external properties such as the labels it is impossible to find out the validity of the clusters. From the problem statement we have the labels so we will be able to check the internal and external cluster validity index to evaluate the clusters. The procedure to cluster the data is elaborated in steps below.

(a) It is understood that we need to cluster the given input data of 300 images into 2 clusters, Male and Female. So K = 2

(b) We run the following steps for different centroid seeds, such as zero vector & first two input data. (The best performance was obtained with zero vector and hence we finalized with this as the seed)

(c) With the seeded mean we populate the membership matrix using the membership function shown in the figure 5. We use euclidean distance as the similarity measure between the vectors $x_i \& v_j$ in every step.

$$u_{j,i} = \begin{cases} 1, & \|x_i - v_j\| \leq \|x_i - v_q\|,\ q \neq j \\ 0, & \text{otherwise} \end{cases}.$$

Fig. 5. Membership Function

$$v_j = \sum_{i=1}^{n} u_{j,i} x_i \Big/ \sum_{q=1}^{n} u_{j,q}$$

Fig. 6. Update Mean Function

(d) After populating the membership matrix with the membership function, we recalculate the mean for the new members as portrayed by the new membership matrix. We use the formula shown in figure 6 to update the mean.

$$\min_{U,V} \left\{ J(U,V) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{j,i} \|x_i - v_j\|^2 \right\}$$

Fig. 7. Objective Function

(e) After a new mean or the new centroid has been calculated the Cost is calculated using the formula in figure 7

(f) We repeat the steps 3 4 5 till J or the cluster centroids/mean does not change

*1) **Centroid Seed Choice**:* In the case mentioned above we choose the mean to be zero vector, but in actual scenario we can let the centroid be any vector in the input sample space. This leads us to understand that the performance of the K means algorithm will vary for different starting centroids. The K means's disadvantage is it minimizes the objective function to reach the local minimum and not the global minimum. Hence the seed for the centroid is an important factor to achieve proper cluster validity. The seed decides the following the maximum number of iterations required to reach the local minimum, hence reaching a poor convergence rate and hence produce sub optimal clusterings Will result in suboptimal final cost. (Final J)

Another known method is to initialize the parameters in a gaussian mixture model. [3]

IV. Evaluations

This section explains the various steps taken to evaluate the systems described above.

PCA

1) With the 200 probe images and 100 gallery images, we calculate the *Similarity Matrix* of size 200x100 where every image from the probe set is compared with the gallery set to get a similarity measure and this similarity measure is used to populate the similarity matrix. Since our similarity measure is euclidean distance the lesser the value the more similar it is.
2) We can obtain the genuine score from the diagonal elements and the rest constitute the imposter scores. The genuine imposter distributions are the probability distributions for the genuine and imposter matches in the dataset. The area shared by both the curves gives a measure of how close the imposter and genuine comparisons are. A system that separates this genuine and imposter curves the max is said to perform better than the others.
3) The cumulative match characteristics curve gives the rank-t identification rate. CMC curve gives us the least rank to implement to get max recognition rate. Thus if this is achieved sooner the better is the system's performance.
4) The receiver operator characteristics curve(roc) is calculated using the false match rate and the (1-false non match rate). The false match rate is the number of imposters that are lesser than the threshold divided by the number of imposter comparisons . The false non match rate is the number of genuine scores greater than the threshold divide by the number of genuine comparisons. The ares under the curve should be close to 100% the higher the area under the curve the better is the performance.
5) From the images 9 10 11 12 13 14 15 16 17 18, we can understand that there is not much change in performance of the system as the number of coefficients increase from 10 to 100. Or the major variance of the dataset is available in the top 10 eigenvectors and hence incorporating them gives us the optimal performance that can be obtained from this input data. But as a whole as there is not much variation in the plots.
6) Even though there was not much overall difference in the above mentioned figures, from figure 8 we can infer that we need at least 60 coefficients for obtaining the best rank1 recognition rate. Rank 1 recognition rate gives the accuracy for the system as a whole when keeping the highest threshold possible. Keeping a higher threshold corresponds to moving the threshold towards the left in the genuine and imposter distribution graphs. Moving the threshold towards the left leads to reducing the number of false matches and increasing the number of true negatives. Increasing the true negatives is going to reduce the accuracy of the system and reducing the false positives is going to increase the accuracy of the system. We have negotiate between these to get the best operating point. For achieving this best optimal operating point the rank 1 recognition rate gives us a measure of how high the threshold can be moved without reducing the accuracy of the system. This curve along with CMC curve can be used to compare different systems.
7) Increasing the number of coefficients beyond this is not going to increase the performance of the system as we can see the recognition rate has leveled after 60 coefficients. This leads us to assume that most of the information is contained in the top 60 eigenvectors.
8) Using euclidean distance as a measure of similarity, we calculate the similarity matrix between the probe and the gallery images directly without reducing the dimensionality. The performance of this system is available in the figure 19. By comparing the figure 19 and the characteristics for the system with PCA, we can clearly see that the area shared under the genuine and imposter is lesser for the system without the PCA, the area under the ROC curve is more for the system without PCA. From this we can infer that the system without PCA performs way better than the system with PCA. In this case using PCA reduces the performance of the system. There can be various reasons for this. Both the gallery and the probe data is perfectly normalized (which can never be achieved in the normal scenario). In case the images are not perfectly normalized we will get small variations in illumination and pose which can be overcome during recognition by using PCA. A similar case is explained in the [2], where they use an approximation procedure to normalize the slight variations in the data.

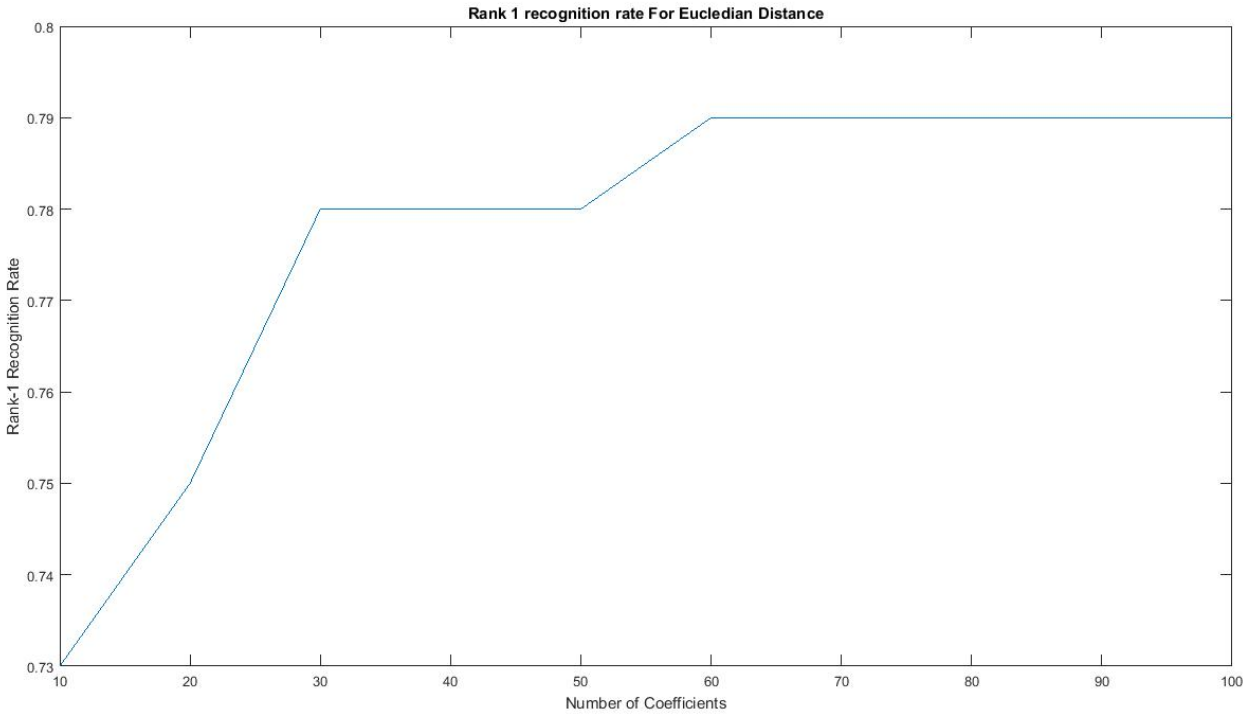

Fig. 8. Rank 1 Recognition Rate vs Number of Coeffecients

K means

1) We run the KMeans algorithm for different centroids. First we try it for the first 2 inputs from the dataset passed to the KMeans algorithm to get an accuracy of 57% percentage. For the centroid seeded with zero mean we get an accuracy of 60.3%. Thus we choose the zero mean as the standard seed for the rest of the experiments.
2) By taking 10 to 100 coefficients(in steps of 10) and euclidean distance as the distance measure we perform the K means algorithm to cluster the input data of total 300 images. The resultant clusters are validated for the different validity indices explained in the later poitns. Since using one criteria to measure the goodness of the cluster is not a good practice, we perform validation on various indices and try to analyze the data.
3) The Internal criteria used for validation are Davies Bouldin Index(DB) and the Calinski-Harabasz(CH) index. The plot of these two indices varying for different numbers of coefficients used is available in the figure 21. The number of coefficients that minimizes the DB index and the number of coefficients that maximizes the CH index is better. From the graph we can clearly get a winner the least number of coefficients show the best validity. But from both the curves we can see that they show a perfect cure either increasing or decreasing. So we check the performance with another index the silhouette index (available in the figure 21). From the three graphs we can find that the rate of change of the index is constant for number of coefficients at 20 and 30 hence leading us to infer that it would be the optimal number of coefficients.
4) The External Criteria used for validation is Entropy and F1measure (fig 23). Both measure almost similar properties of the cluster, true positives, true negatives, false positives and false negatives. From the figure 23 we can see that the entropies are not constant and we repeat the step with K mean seeded with zero vector to obtain figure 25. From the this fig 25, we can see that the validity remains constant for the various number of coefficients. This leads us to understand that the PCA does not have any effect on the clustering. The accuracy of the system also remains around 60.30%. This proves that the PCA has no effect on the clustering. This was clearly not visible in the internal criteria because the internal criteria did not take into account the actual labels of the data. In our case we have the external validity criteria(labels of the data) and hence a more exact measure of the goodness is calculated in this case.
5) Clustering the data without applying PCA also results in an approximately same accuracy 60%. Thus we can confirm that PCA does affect clustering of the data for soft biometrics.
6) This trend is not the same as in the recognition section. We can clearly see that increasing the components in-

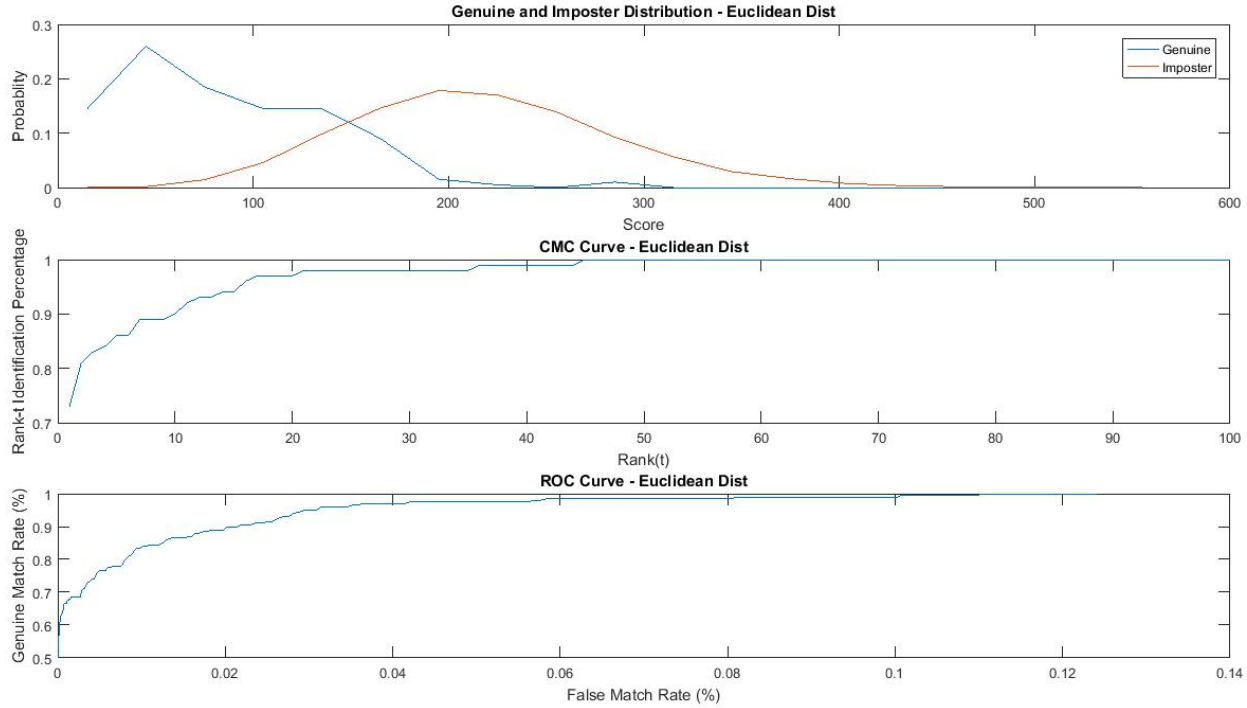

Fig. 9. The Performance Characteristics for top 10 Coefficients

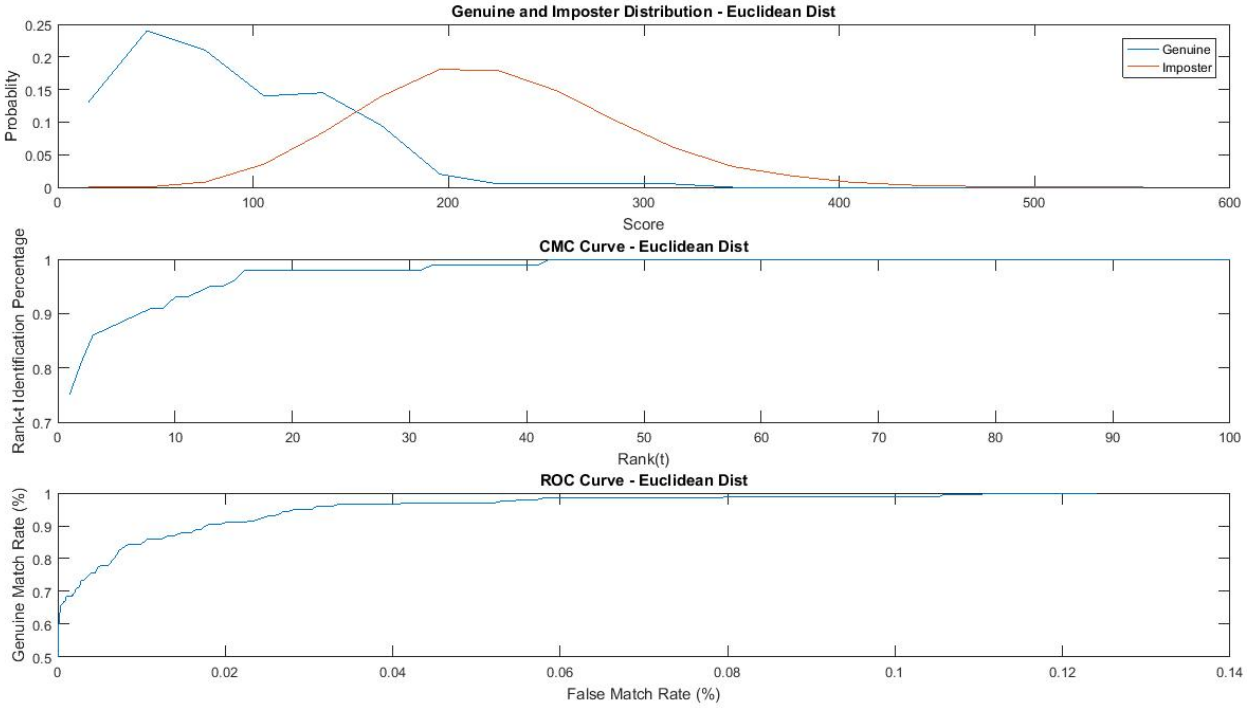

Fig. 10. The Performance Characteristics for top 20 Coefficients

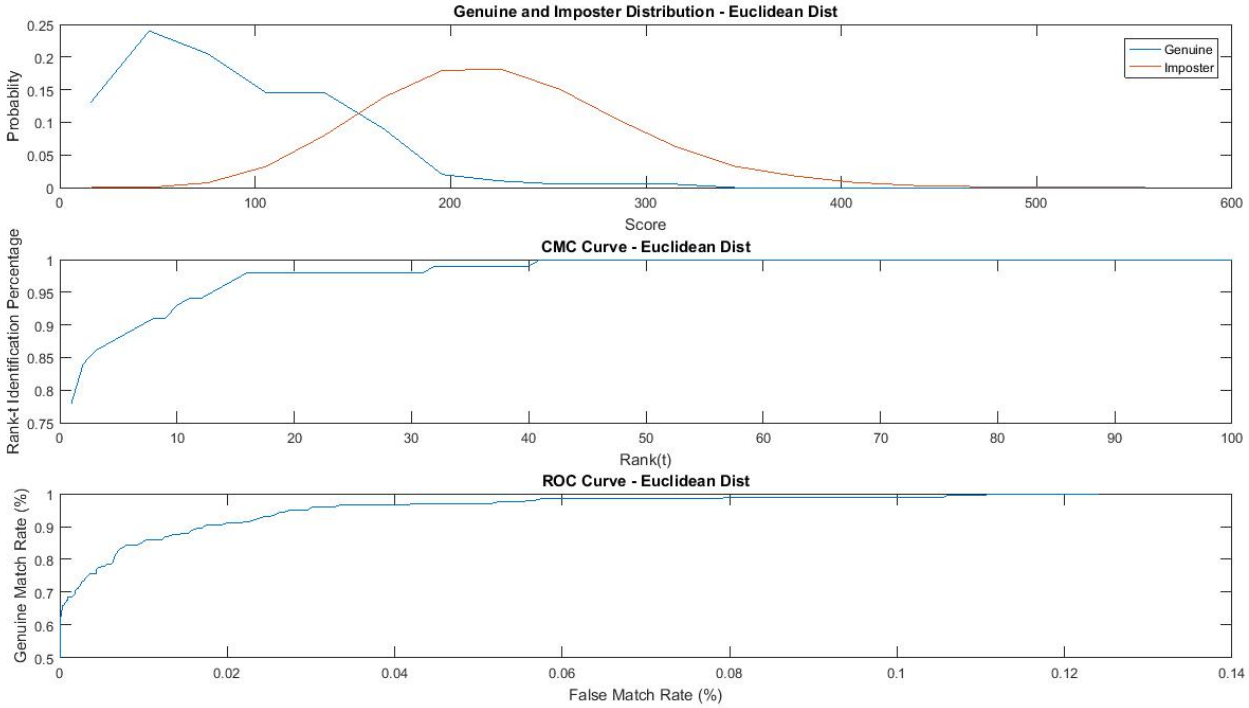

Fig. 11. The Performance Characteristics for top 30 Coefficients

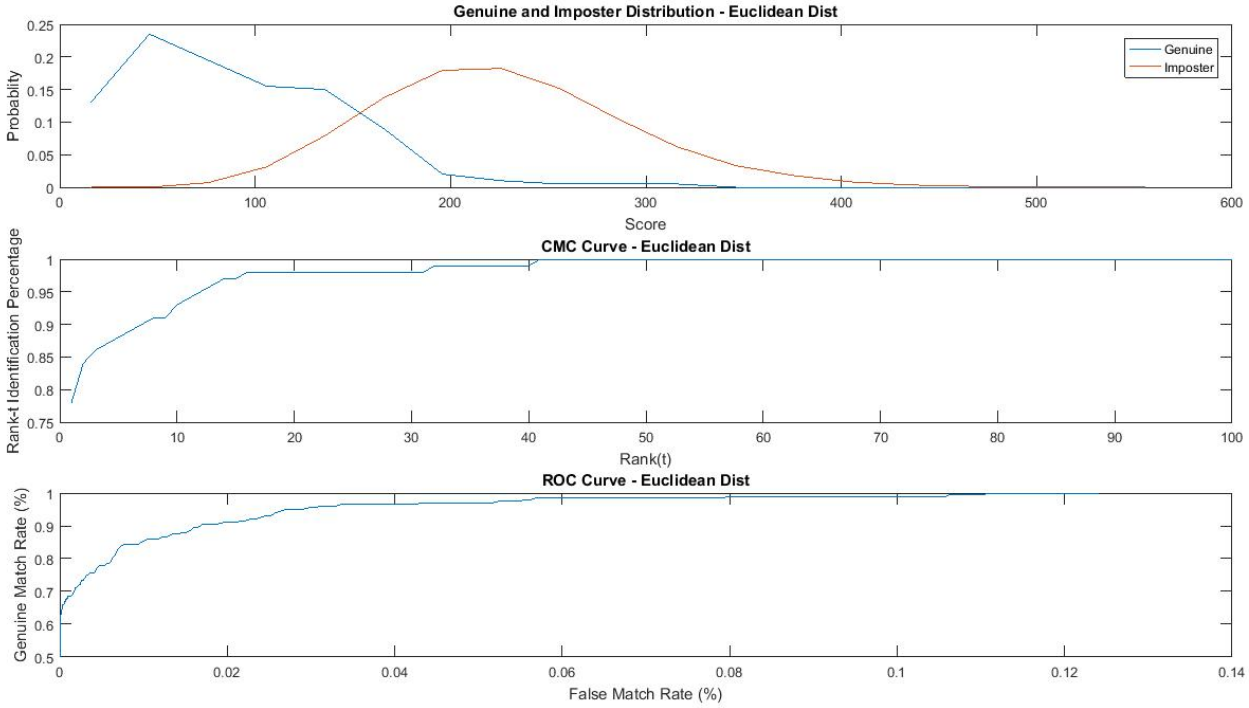

Fig. 12. The Performance Characteristics for top 40 Coefficients

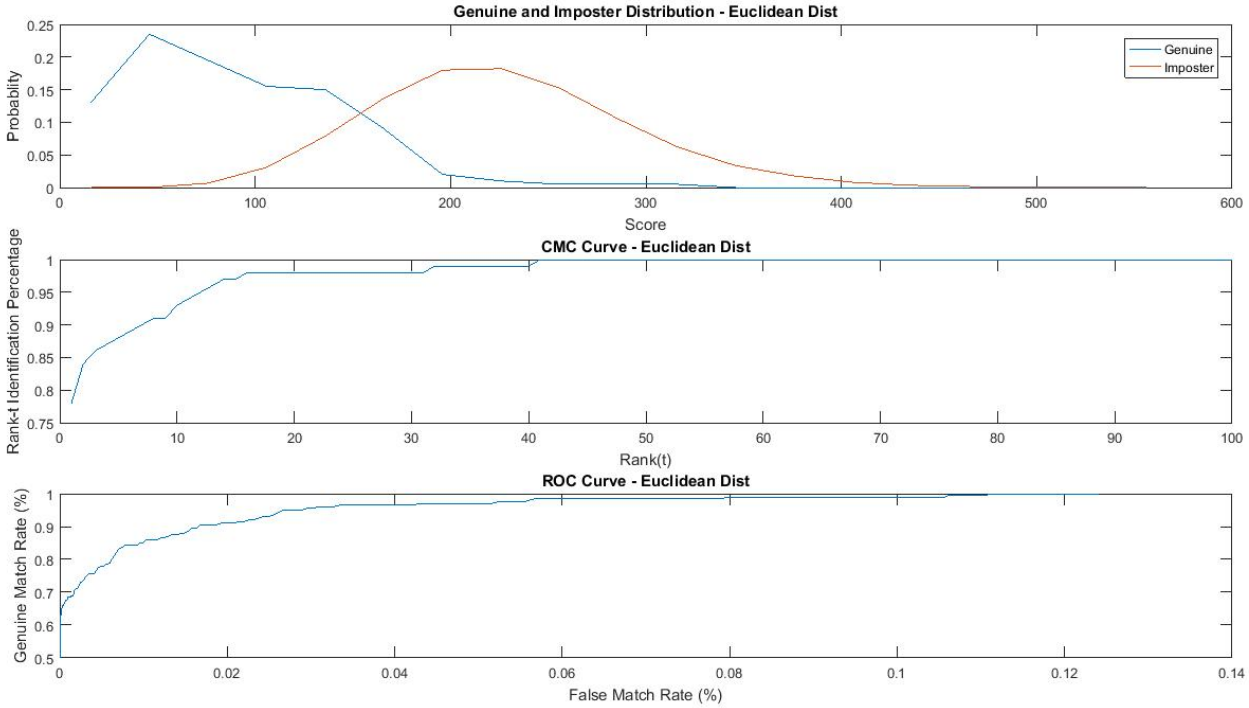

Fig. 13. The Performance Characteristics for top 50 Coefficients

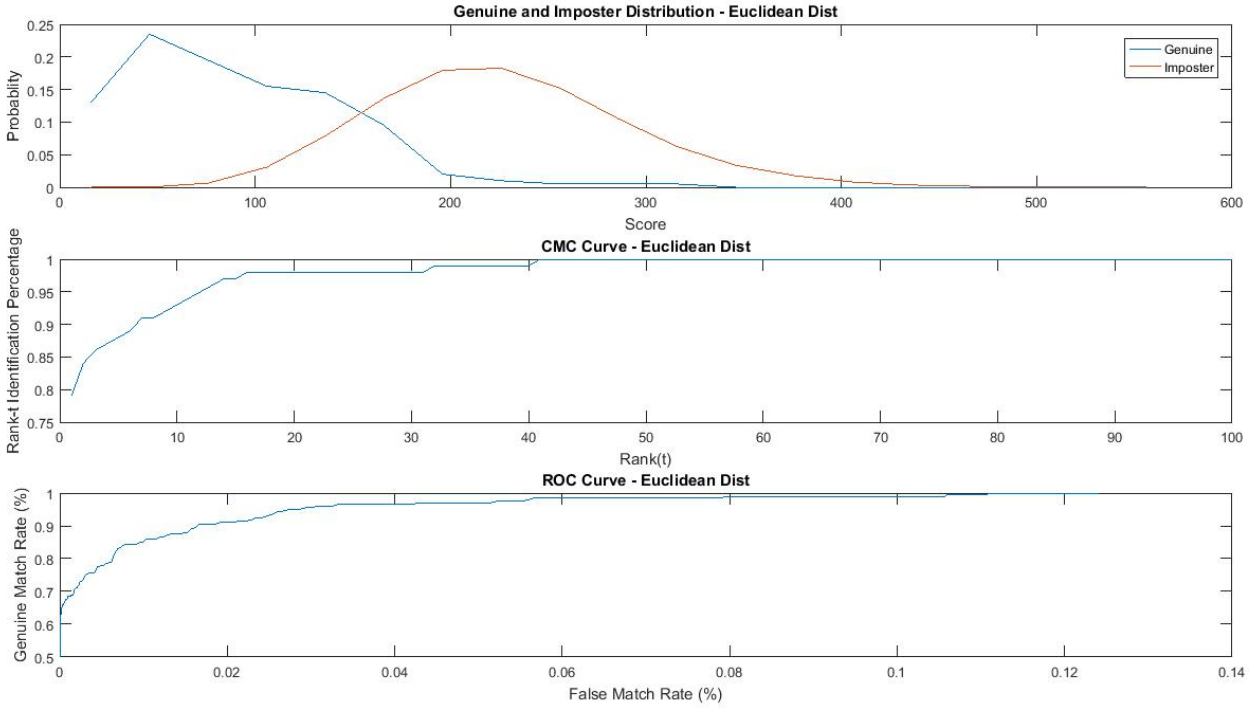

Fig. 14. The Performance Characteristics for top 60 Coefficients

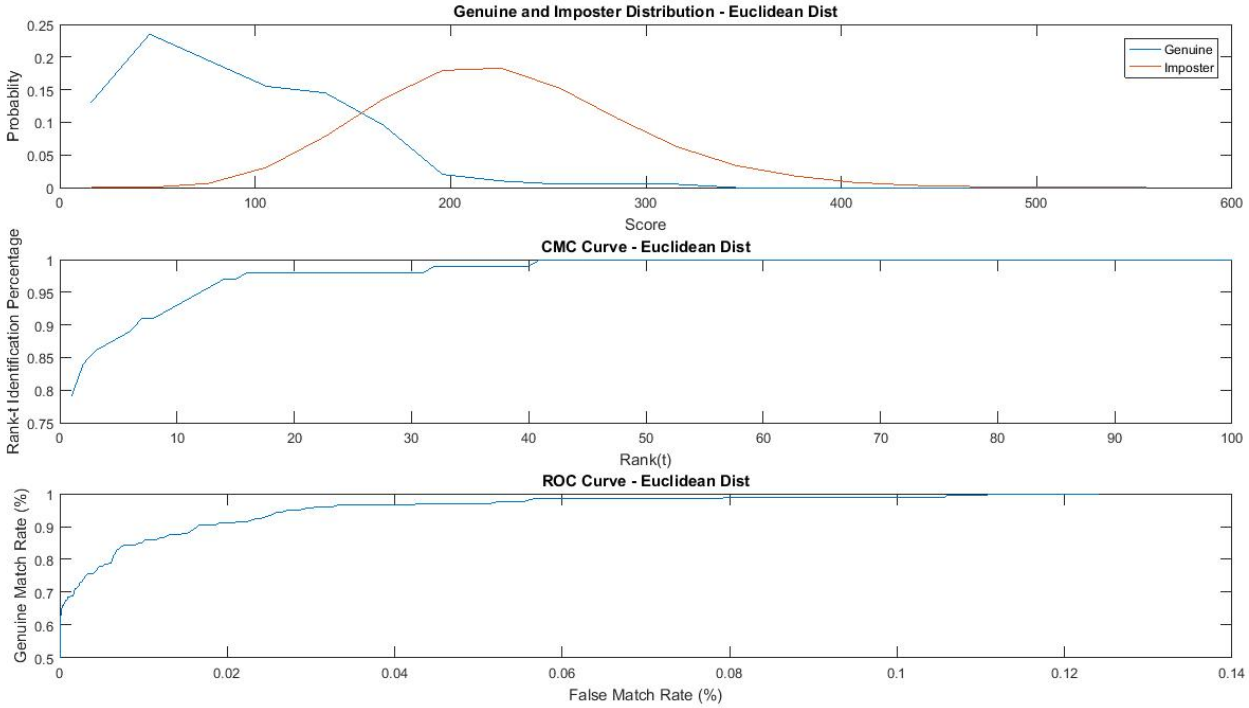

Fig. 15. The Performance Characteristics for top 70 Coefficients

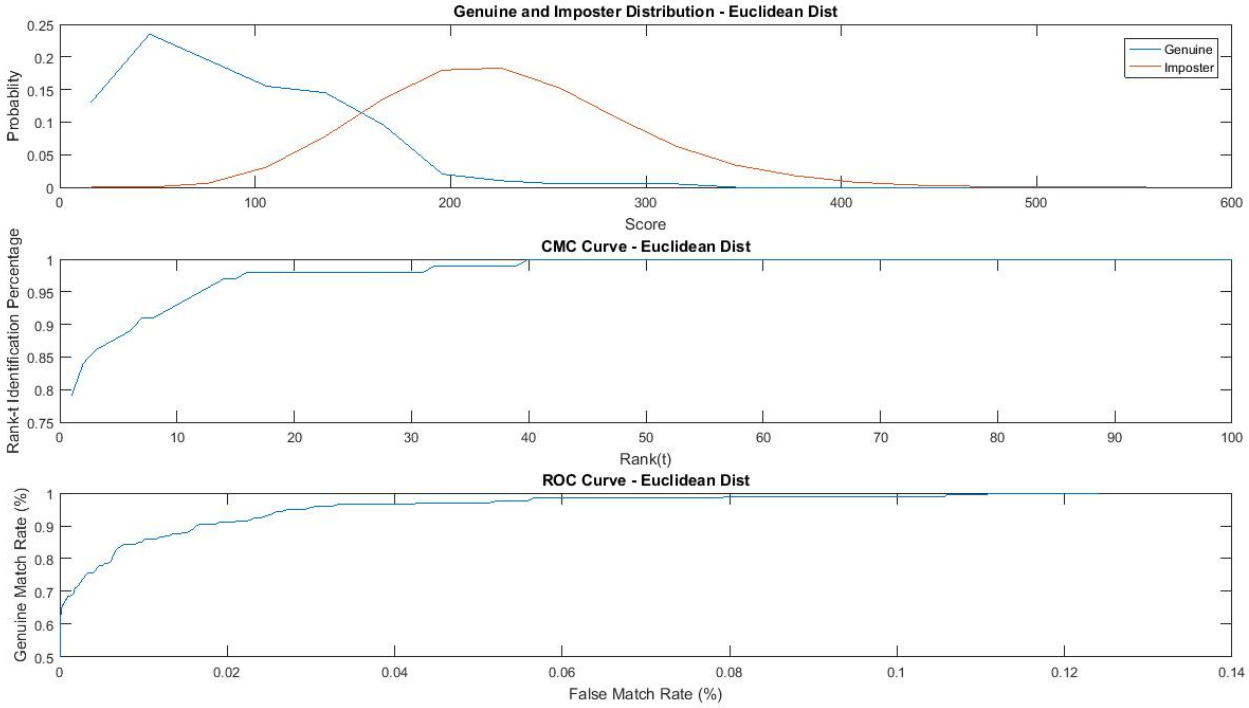

Fig. 16. The Performance Characteristics for top 80 Coefficients

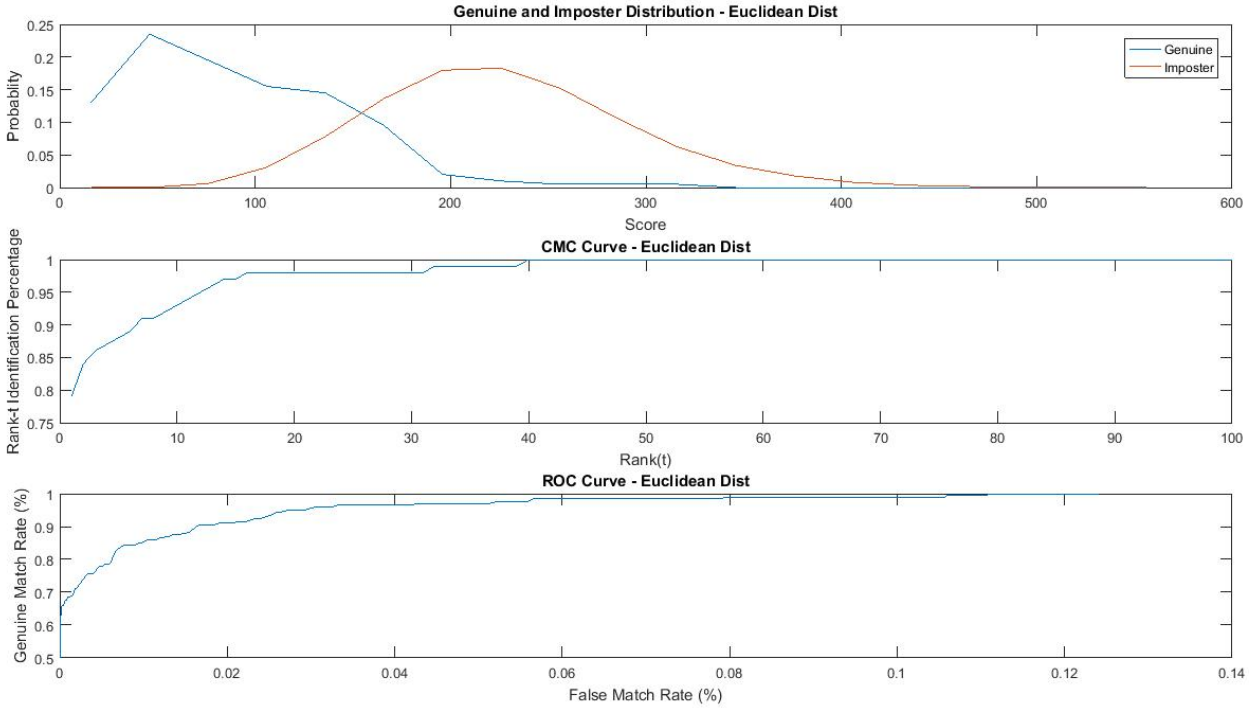

Fig. 17. The Performance Characteristics for top 90 Coefficients

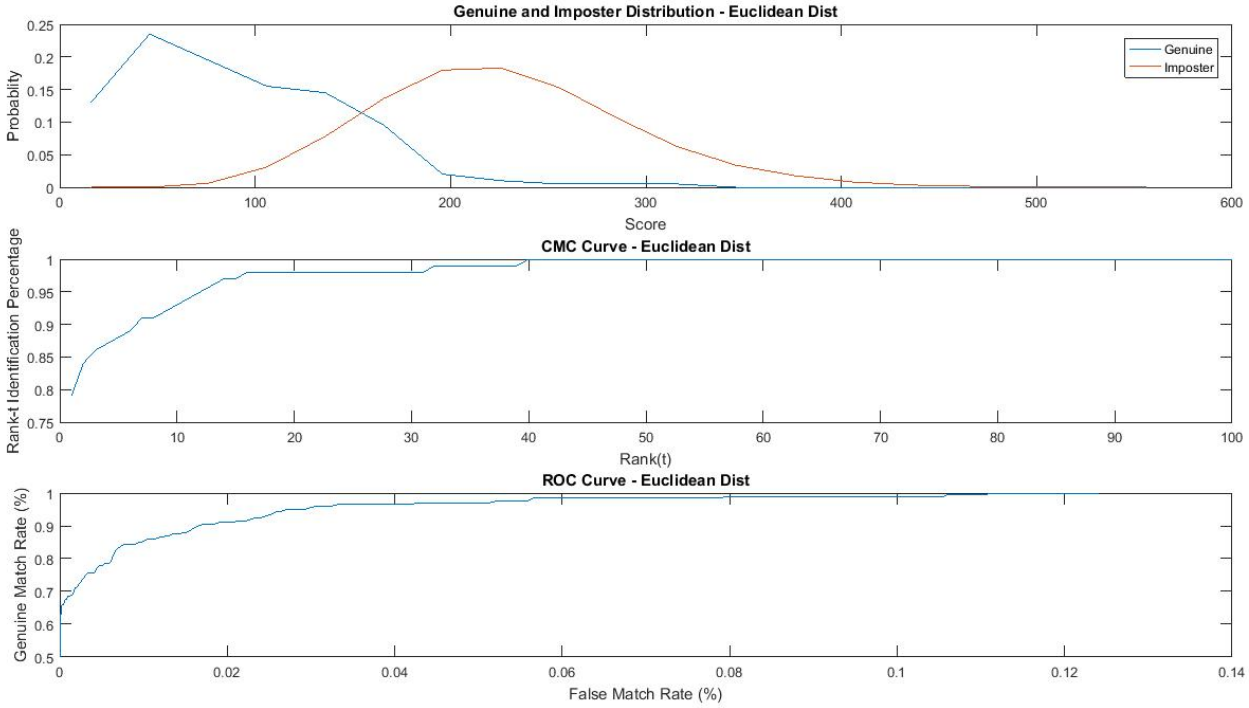

Fig. 18. The Performance Characteristics for top 100 Coefficients

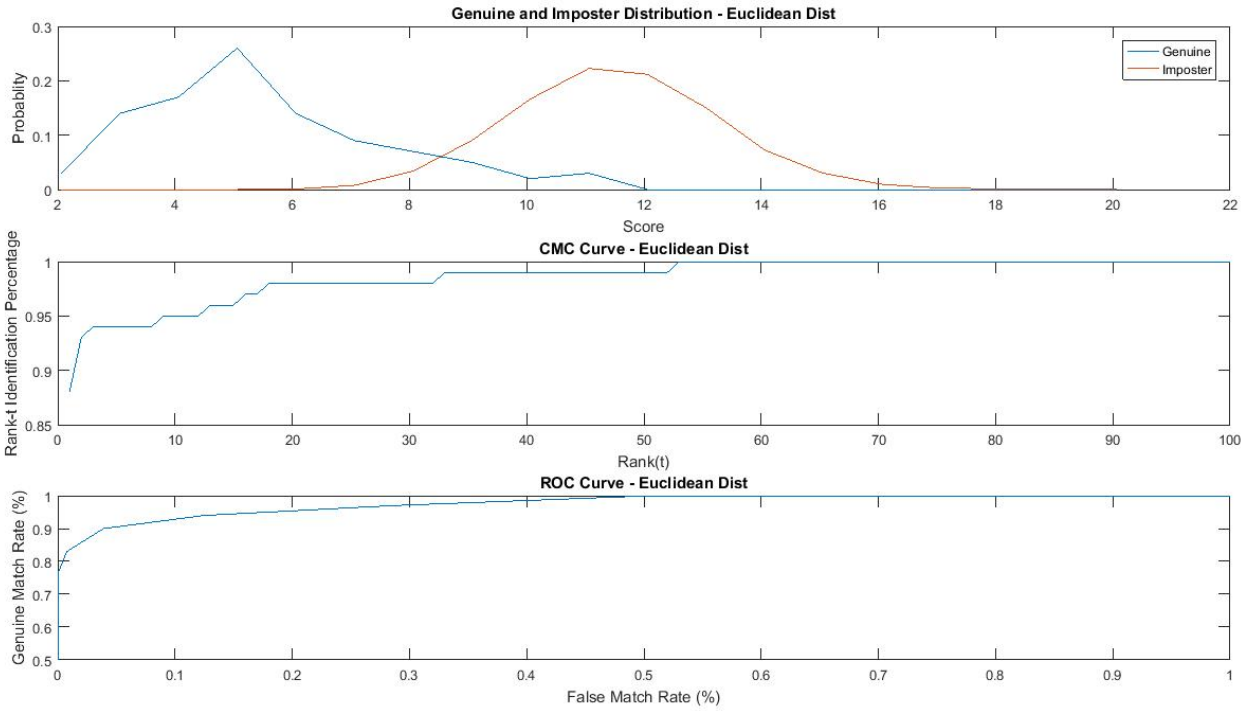

Fig. 19. The performance Characteristics without PCA

crease the rank1 recognition rate. But increasing the components for clustering does not change the performance.

V. Conclusion

1) The top 3 eigen faces are available in figure 4. The first principal component in terms of facial component corresponds to illumination invariance.
2) The recognitionperformance does not vary after an initial increase in performance
3) The PCA does not have an effect on the clustering.

References

[1] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 4-20, Jan. 2004. doi: 10.1109/TCSVT.2003.818349

[2] Sirovich and Kirby "Low Dimensonal Procedure for Characterisation of Human Faces" (1987)

[3] Bishop, Christopher M. "Pattern recognition." Machine Learning 128 (2006). APA

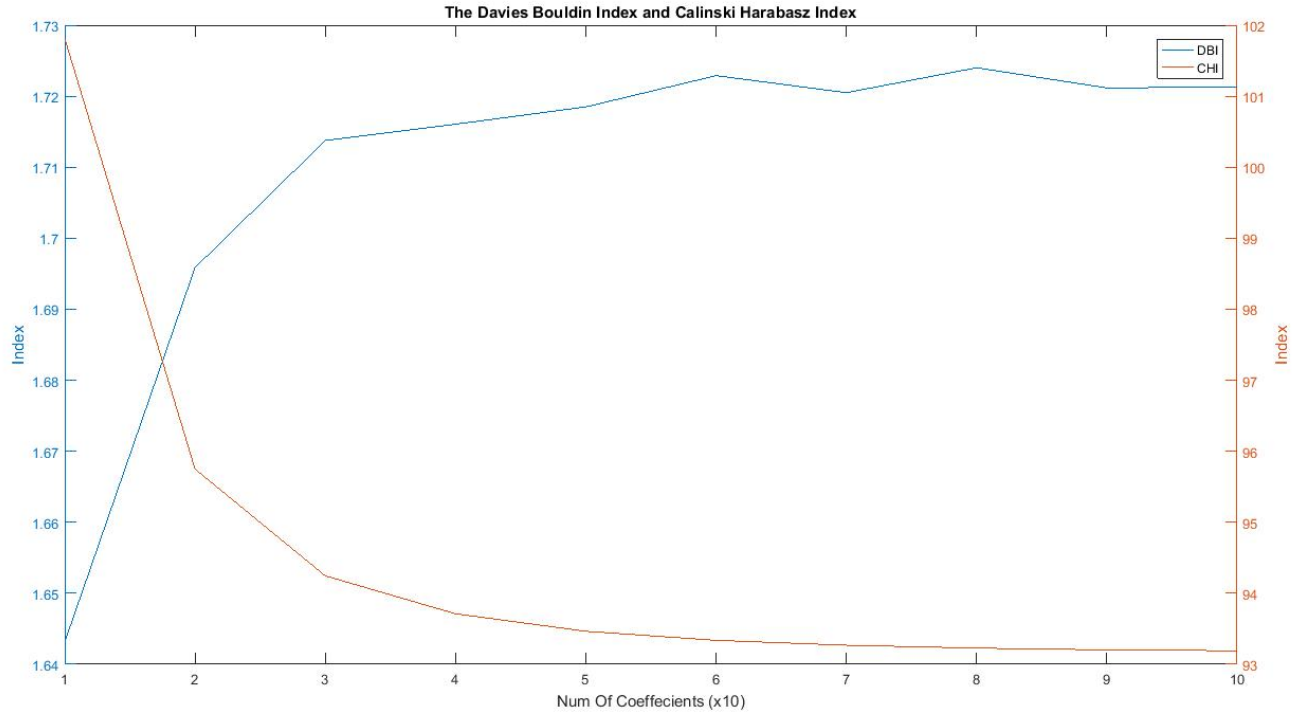

Fig. 20. Internal Validity Criteria vs Number of Coefficients

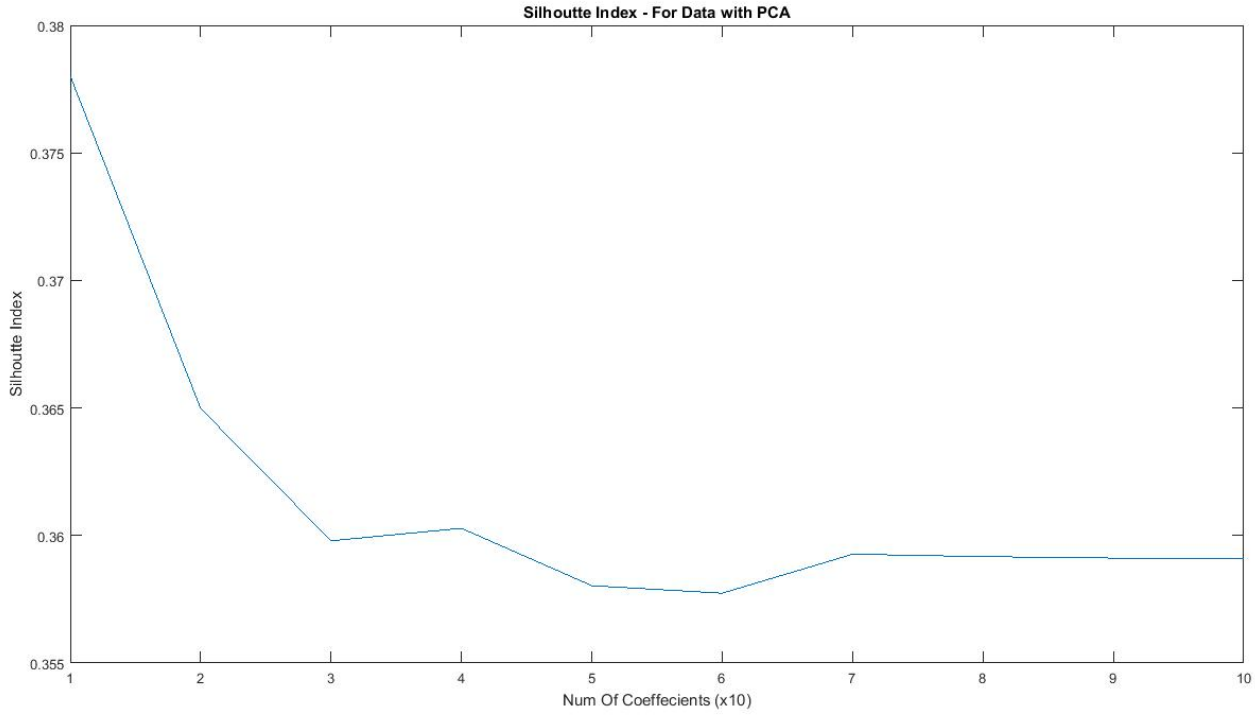

Fig. 21. Silhouette Index vs Number of Ceofficients

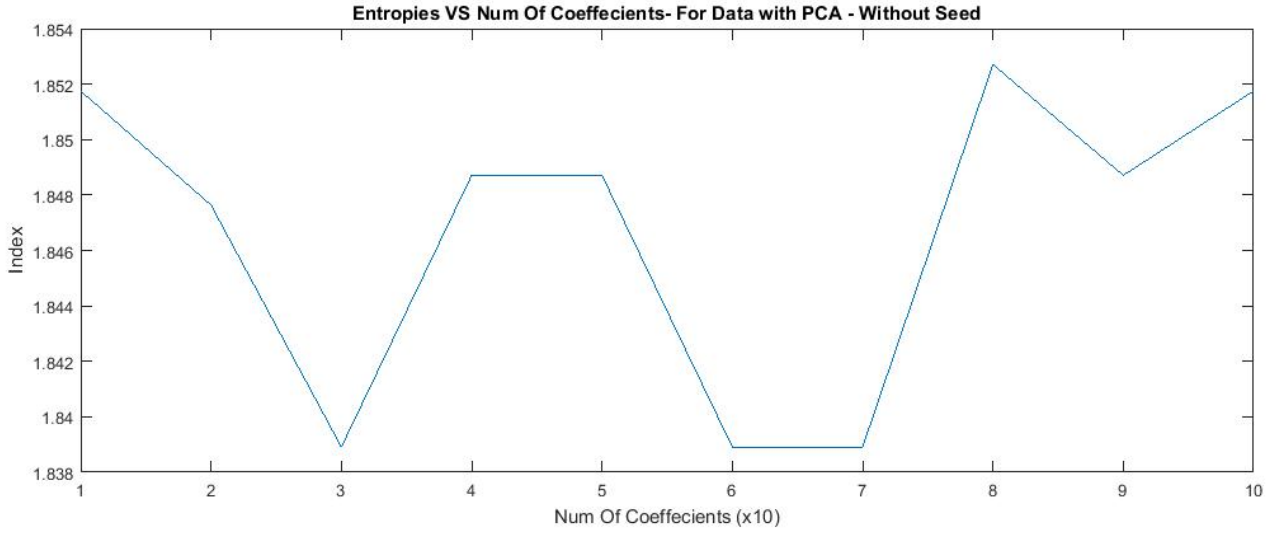

Fig. 22. Entropy Validity Criteria - unseeded K mean

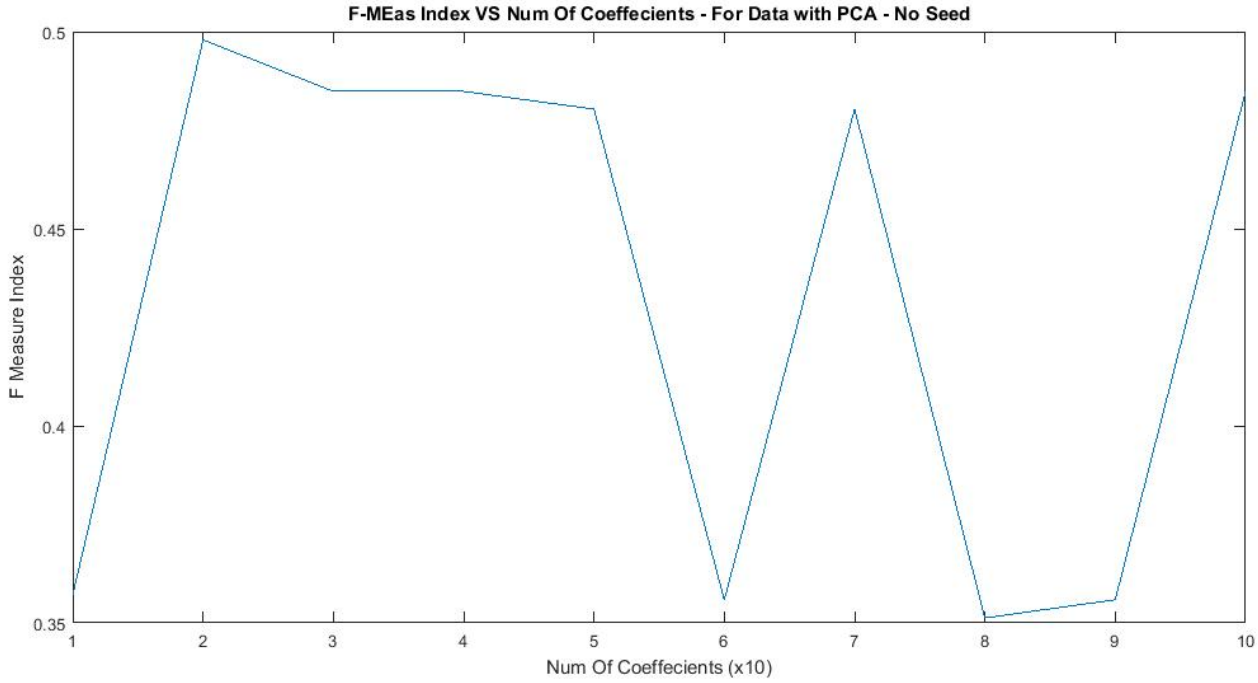

Fig. 23. F1Measure Validity Criteria - unseeded K mean

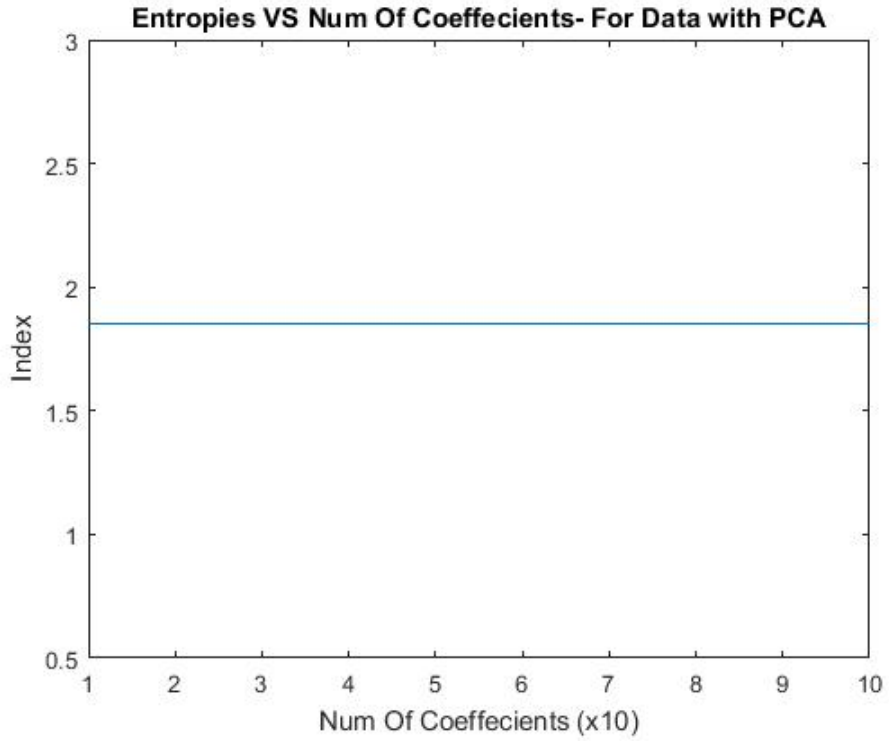

Fig. 24. Entropy Validity Criteria - K Mean seeded with zero vector

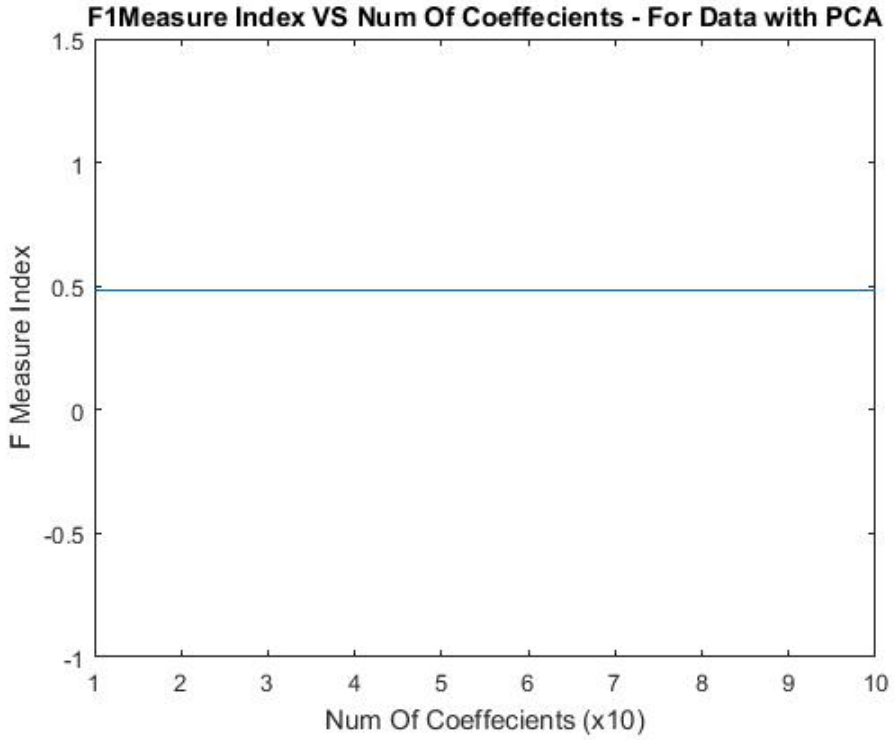

Fig. 25. F1Measure Validity Criteria - K Mean seeded with zero vector