# DIMENSIONALITY'S BLESSING

## CLUSTERING IMAGES BY UNDERLYING DISTRIBUTION

CVPR 2018

EE798R COURSE PROJECT


AUTHORS

Wen-Yan Lin, Siying Liu, Jian-Huang Lai, Yasuyuki Matsushita

INSTRUCTOR: DR. TUSHAR SANDHAN
MENTOR: HIDANGMAYUM BEBINA DEVI
BY: VISHAL HIMMATSINGHKA (211175)


## INTRODUCTION

In high-dimensional data analysis, traditional clustering approaches often struggle due to the **"contrast-loss"** phenomenon, where distances between data points converge, making separation challenging. This paper introduces a fresh perspective, viewing "contrast-loss" not as a hindrance but as an opportunity. By leveraging this phenomenon, the authors demonstrate that data points from the same distribution tend to concentrate on thin, distinct hyper-shells, facilitating separability even in overlapping datasets.This novel insight transforms contrast-loss from a limitation into a powerful tool for data organization.

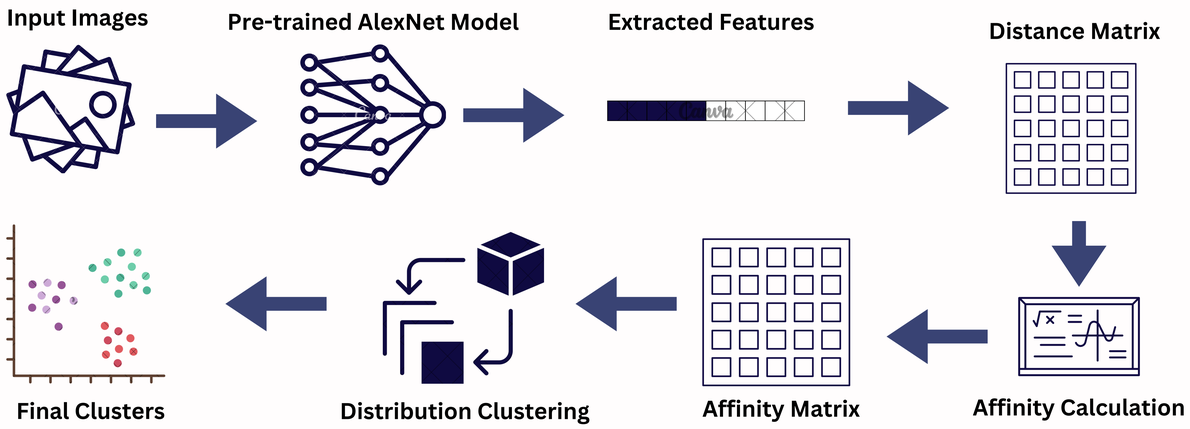
## OBJECTIVE

This research aims to develop an innovative clustering algorithm, termed "**Distribution-Clustering**," that exploits contrast-loss in high-dimensional spaces. Unlike conventional methods, this technique automatically determines the number of clusters, effectively groups data by underlying distributions, and exhibits a strong robustness to outliers. The approach redefines clustering accuracy, especially for chaotic, high-dimensional datasets, by grouping data based on statistical distribution similarities.

## RELATED RESEARCH

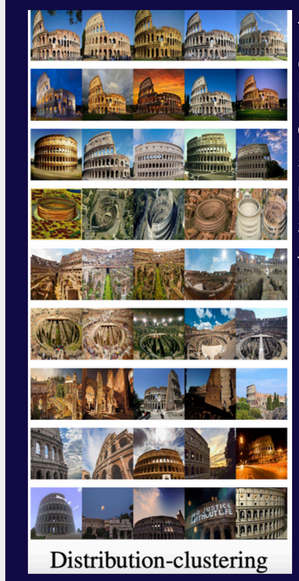
The challenges posed by high-dimensional clustering have led to various methods, including sub-space clustering and projective clustering, designed to mitigate contrast-loss. However, recent studies suggest that contrast-loss could offer advantages in cluster and outlier detection. Building on foundational work by Beyer et al., which explored dimensionality's impact on clustering, this paper extends those findings by demonstrating that contrast-loss can reveal natural data separability, forming a basis for the proposed clustering algorithm.

## METHODOLOGY




1. **Feature Extraction**: A pre-trained **AlexNet model** in the TestModel class extracts deep features for each image, converting them into one-dimensional vectors through adaptive pooling and flattening to capture essential characteristics.
2. **Distance Calculation**: The cluster function computes pairwise squared Euclidean distances between image feature vectors, laying the foundation for clustering by measuring the "**closeness**" of images in feature space.
3. **Affinity Matrix Computation**: Using these distances, the compute_affinity_matrix function calculates a second-order affinity matrix. This matrix reflects image similarities by computing and averaging the squared differences in distances of image pairs $i$ and $j$ to all other images, capturing second-order similarity in elements $\mathbf{A(i, j)}$.
4. **Distribution-Clustering Algorithm**: The algorithm initializes clustering with the smallest non-zero element in the affinity matrix, forming a candidate cluster $H$. It expands $H$ by including points with average second-order distances below a threshold $\tau$, and validates clusters that meet a minimum size, assigning unique labels.
5. **Outlier Handling**: Images that don't fit well into existing clusters are treated as outliers and placed separately, preserving the purity of primary clusters.
6. **Final Clustering Result**: This process repeats until all images are either clustered or marked as single-point outliers if they don't meet similarity criteria, resulting in clusters that group images based on their underlying distribution characteristics within the threshold tau.

## RESULTS




The clusttering results of Distribution clustering was compared with other known methods like **K-Means, GMM, Spectral clustering** on various performance metrices such tas **Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index**. Some sample results can be seen below on two given datasets, coloseum images and Mnist digit dataset are shown below.

| method | Silhouette Score | Calinski-Harabasz Index | Davies-Bouldin Index |
|---|---|---|---|
| kmeans | 0.025718 | 4.999815 | 2.248301 |
| gmm | 0.021279 | 4.790093 | 1.896970 |
| spectral | 0.025493 | 4.747467 | 2.313330 |
| distribution | 0.028362 | 3.192529 | 2.287611 |
| improved_distribution | 0.034347 | 3.462428 | 2.224356 |

| method | Silhouette Score | Calinski-Harabasz Index | Davies-Bouldin Index |
|---|---|---|---|
| kmeans | 0.088517 | 27.638248 | 2.217181 |
| gmm | 0.100902 | 27.930472 | 2.153142 |
| spectral | 0.070360 | 26.062310 | 2.231486 |
| distribution | 0.189591 | 17.558190 | 1.549544 |
| improved_distribution | 0.186540 | 18.402878 | 1.805982 |

## PROPOSED ADVANCEMENT

The proposed enhancement to the Distribution-Clustering algorithm introduces **Fisher Linear Discriminant (FLD)** as a substitute for Euclidean distances in constructing the distance matrix. Starting from the initial clusters obtained through some method, FLD projects the data into a space that maximizes inter-cluster separability while minimizing intra-cluster variance. This transformation yields a refined distance measure, allowing the affinity matrix to more accurately represent similarities between projected vectors

## DATASET

Tests had been performed on a given unit dataset of 100 images alongwith the mentioned known datsets such as MNIST and Caltech 101 with multiple sets of images.

## CONCLUSION

In conclusion, this paper presents a novel perspective on the contrast-loss phenomenon, traditionally seen as an obstacle in high-dimensional clustering. By reinterpreting contrast-loss as a beneficial property, the Distribution-Clustering algorithm leverages this effect to group data points based on underlying distribution similarities, producing clean clusters with automatic outlier handling. Additionally, the proposed enhancement introduces the Fisher Linear Discriminant (FLD) method to refine the clustering process further. By employing FLD in constructing the affinity matrix, the algorithm aims to improve cluster separability and robustness, especially in complex, high-dimensional data scenarios. This approach not only enhances clustering accuracy but also extends the algorithm's applicability to more challenging datasets.