symmetry

MDPI

Article

Probabilistic Unsupervised Machine Learning Approach for a Similar Image Recommender System for E-Commerce

Ssvr Kumar Addagarla and Anthoniraj Amalanathan *

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India; ssvrkumar.addagarla2017@vitstudent.ac.in

* Correspondence: aanthoniraj@vit.ac.in

Received: 28 September 2020; Accepted: 23 October 2020; Published: 27 October 2020

Abstract: The recommender system is the most profound research area for e-commerce product recommendations. Currently, many e-commerce platforms use a text-based product search, which has limitations to fetch the most similar products. An image-based similarity search for recommendations had considerable gains in popularity for many areas, especially for the e-commerce platforms giving a better visual search experience by the users. In our research work, we proposed a machine-learning-based approach for a similar image-based recommender system. We applied a dimensionality reduction technique using Principal Component Analysis (PCA) through Singular Value Decomposition (SVD) for transforming the extracted features into lower-dimensional space. Further, we applied the K-Means++ clustering approach for the possible cluster identification for a similar group of products. Later, we computed the Manhattan distance measure for the input image to the target clusters set for fetching the top-N similar products with low distance measure. We compared our approach with five different unsupervised clustering algorithms, namely Minibatch, K-Mediod, Agglomerative, Brich, and the Gaussian Mixture Model (GMM), and used the 40,000 fashion product image dataset from the Kaggle web platform for the product recommendation process. We computed various cluster performance metrics on K-means++ and achieved a Silhouette Coefficient (SC) of 0.1414, a Calinski-Harabasz (CH) index score of 669.4, and a Davies–Bouldin (DB) index score of 1.8538. Finally, our proposed PCA-SVD transformed K-mean++ approach showed superior performance compared to the other five clustering approaches for similar image product recommendations.

Keywords: PCA-SVD dimensionality reduction; K-means++ clustering; similar image recommender system; Manhattan distance; cluster similarity

1. Introduction

In the digital era, the e-commerce industry is growing rapidly, especially since many of the users are migrating towards online shopping from traditional offline shopping in many developing countries. In countries such as India, China, Singapore, Malaysia, Japan, etc., the growth rate consistently increases, and millions of users are interested in purchases through e-commerce platforms [1,2]. There is a wider scope of increase in the demand for e-commerce purchases in many developing nations due to the COVID-19 pandemic situation, and millions of people prefer to purchase through online shopping now-a-days. Especially several categories of products are in high demand like clothing, electronics, furniture, sports, etc. [2]. Out of which apparels and some of the electronic products mostly rely on visual appearance to attract the users to purchase the products. These e-commerce portals consist of millions of images relevant to various products, and bringing the desired product of the customer is a challenging issue. Many researchers had come up with several possibilities to address

the problem but not a satisfactory solution to many e-commerce problems [3–6] as many e-commerce platforms use a recommender system and mainly rely on text-based searching approaches. It takes the user input and bases it on the word tokenization process. The recommender system is to bring the possible matching products to the users. However, this kind of approach had some limitations and missed many features such as colors, patterns, texture, and shape in the product images [7,8]. In this text-based search, there is a need to specify the descriptions of each product, which may not describe the entire product. Here, we come up with a possible solution, which is to search for the relevant product bases on the given input image, as shown in Figure 1. We try to train the model for the various features through the feature engineering process, such as colors, texture, and the shape of the objects in the image, which further can be processed through various machine learning dimensionality reduction techniques, such as Principal Component Analysis (PCA) [9], Expectation-Maximization (EM) PCA [10], Linear Discriminant Analysis (LDA) [11], and Probabilistic PCA [12,13], for the proper recommendations from the given input image [14]. Once the features are extracted, it is necessary to compute the similarity measure in terms of distance to be computed from the origin to the set of images in the database [15].

Figure 1. Outcome of the proposed final recommender model.

Several clustering-based unsupervised learning approaches have been proposed by researchers for a similar image recommender system where the ground truth labels are known. For measuring the distance and analyzing the quality of the recommender model, the following measures [16] are to be used for the unsupervised learning approach.

- Adjusted Rand Index (ARI) is used to measure the similarity between the clusters where ground truth labels are known [17]. The ARI can be computed using Equations (1) and (2) [16].

$$ARI = \frac{Random\ Index - Expected\ Random\ Index}{Maximum\ Random\ Index - Expected\ Random\ Index} \quad (1)$$

$$Random\ Index = \frac{x + y}{S_2^{n_{samples}}} \quad (2)$$

where x, y are the number of elements with the same and different set in S of the cluster.

- Homogeneity (hg), Completeness (cp) are the measures for the same class and same cluster predictions [18], and harmonic mean can be computed using V-measure [19] as described in Equation (3) [16].

$$V_{measure} = \frac{(1+\beta) \times hg \times cp}{(\beta \times hg + cp)} \tag{3}$$

where β is the random value to be randomly considered as less than 1, the harmonic mean value ranges between 0 and 1, and the highest value will be the best to consider.

- The Fowlkes–Mallows score (FMS) [20,21] is used to compute the geometric mean of the similarity of the clusters where the ground truth labels are known where the FMS is lying between the 0 and 1, and a greater value is a better similarity among the clusters. The FMS uses the True Positives (TP), False Positive (FP), and False Negative (FN) for the similarity measure analysis is shown in Equation (4) [16].

$$FMS = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{4}$$

In our research work, we proposed of approach through unsupervised clustering algorithms where the ground truth labels are unknown. This kind of clustering problem can be computed with performance measures such as the Silhouette Coefficient (SC), the Calinski-Harabasz (CH) index, and the Davies–Bouldin (DB) index, and these measures are discussed briefly in the results section in the paper.

Further, we organized this research work into five major sections, where related works are discussed in Section 2, our proposed method of a similar image recommender system is presented in Section 3, dataset, pre-processing stage, and various experimental approaches are presented in Section 4, the final results and comparisons are presented in Section 5 and the conclusion and limitations of this research work are presented in Section 6.

2. Related Works

Image-based recommendations through clustering play a vital role in many application areas, and profound research works are done mostly on the medical images for various analyses using artificial intelligence and other machine learning techniques. Hancer et al. [22] proposed an artificial bee colony approach for the three benchmarks Lena, Remote Sensing, and Brain MRI images for the clustering approach compared with the particle swarm optimization, and K-means cluster approaches. M. Gong et al. [23] had come with a new clustering approach using the Kernal metric for the fuzzy c-mean clustering process and tested on the various image datasets like synthetic, natural, and medical images for the performance evaluation. Karthikeyan and Aruna [24] proposed a probabilistic text and image-based semi-supervised clustering approach. They used the topic modeling, comparing image features such as color sets and image block signature computing the similarity distance measure. Their method further compared with the K-means and Dbscan unsupervised clustering algorithms.

Matrix-factorization-based image clustering was proposed by K. Zeng et al. [25], and the objective of their work to recognize the basis on the parts of the objects in the images which are computed through the non-negative matrix factorization technique with the hypergraph regularization process. The authors used the USPS handwritten digit dataset, ORL, and Yale, which are the grayscaled face image datasets used for the performance analysis, and compared the results with pre-existence clustering algorithms such as graph regularized Non-negative matrix factorization (NMF), PCA, and K-means. An improved Artificial Bee Colony (ABC) clustering approach had been tested on seven benchmark images compared to the fitness values, objective functions along with quality measures such as DBI and XBI values with particle swarm optimization, genetic algorithms, and K-means clustering [26].

Younus et al. [27] proposed a content-based similar image retrieval system using Particle swarm optimization (PSO) and K-means by extracting various features in the color images like histograms, co-occurrence matrix, both the color and wavelet moment for the clustering process. For their experimentation, the Wang dataset is used, which consists of 10 label classes with every 100 images. Precision and recall measures are computed on the query and retrieval images and listed out the comparison results with existing methods. The hierarchical and flat clustering approach used to cluster for the cell phone images was proposed [28,29] using sensor pattern noise. In this approach, a total of 1350 images are used to process and cropped and converted to grayscale for applying wiener filter and applied the hierarchical and flat clustering approach. Then, to validate the clustering process, a silhouette coefficient is computed and further applied true positive rate to assess the degree of certainty of the clusters.

Convolutional neural networks have been used to train and perform image clustering by continuously applying the forward and backward propagation to improve the clustering process for representation learning, object loss function, and interpreted through agglomerative clustering process [30]. The work is extensively tested on the benchmark multiple handwritten image datasets and facial image datasets and made a performance comparison with several other algorithmic approaches by using NMI metric. Pandey and Khanna [31] had proposed the Content-based image retrieval (CBIR) approach using agglomerative clustering and applied on labeled multiple datasets and computes various cluster measure for similarity. Vantage Point (VP) trees are a popular approach for the faster indexing of the images in a CBIR approach and in [32] come up with a new distance index measure called DCIVI for the nearest and furthest neighbors by applying the VP-Tree concept for faster image retrieval on a remote sensing image dataset compared with a sequential scanning algorithm in terms of indexing, feature extraction and query response time.

Biradar and Ahmed [33] developed visual CBIR using edge and corner detection of the image using the Harris corner detector and feature extraction of the images done using the Voronoi tree algorithm. Further authors using a support vector machine (SVM) as a classifier of final image classification achieved 90% accuracy on their approach. In [13], the PCA-based unsupervised learning approach was implemented for the segmentation of brain tumor cells from T1 weighted MRI scan images. To this process, authors tested various PCA dimensionality reduction techniques and found that Expectation-Maximization (EM) PCA and Probabilistic PCA (PPCA) are the best fit for their process through FCM clustering and K-means clustering with various image dimensions. A graph-based probabilistic dimensionality reduction using manifold hashing techniques was proposed for image similarity search, which uses the K-nearest neighbor approach for landmark representation and further analysis [34]. Authors had tested their model on benchmark datasets such as CIFAR, MNIST, NUS-WIDE, and GIST and compared with other popular hash methods, such as LSH, ITQ, MDSH, AGH, etc., and shown improvement using their proposed method.

A similar fruit query system was proposed by Fachrurrozi et al. [35] using fuzzy-based feature extraction using a color histogram, and the color moment invariant approach then applied the K-means clustering for the extracted features to compute the similarity distance between the query image and the cluster images done using the K-nearest neighbor search algorithm. In this process, the authors used various single and multi-object fruits to test the accuracy of their model and achieved 92.5% and 90%, respectively. Many researchers have been using converting RGB (Red, Green, and Blue) images to greyscale for feature extraction and further apply various approaches for similar search recommendations. Instead of converting into greyscale, PCA with a smaller dimension of the images can be utilized. Fleece fabric-based image classification has developed using PCA dimensionality reduction and further applies K-nearest neighbors and Naive Bayes classifiers for the accuracy comparison [36]. Various textile image retrieval was proposed using local PCA-based features descriptors like color, orientation for localized color feature extraction and classification done on the join feature criterion for better image retrieval and experimentation carried out on various textile images with different patterns and measured the precision and recall [37].

Chen et al. [38] proposed a novel approach to high dimensional large scale datasets for image classification and retrieval using the local neighboring approach and clustering through NQ-Dbscan and achieves the $O(n * \log(n))$ and $O(n)$ for some cases. Singh and Srivastava [39] had come up with an improved image classifier using Random Forest classifier and LBP feature extraction for text, color, moment, and histogram features. Benchmark Wang dataset is used for performance analysis in terms of accuracy, precision-measure, and MCC are compared with K-NN, SVM, Naive Bayes with the proposed approach. In [40], an artificial neural-network-based approach, along with various image descriptors for color, edge detection using the YCbCr method and discrete wavelet transformation, was applied for better image retrieval, and experiments have been carried out on the Wang dataset and computed the accuracy measures. Later, Recall measures were compared with the existing methods. Jian et al. [41] used the direction patch extraction approach based on the perception and color histogram, and the Gabor filter for texture feature, employed for its process to avoid the complex segmentation, further uses Dbscan clustering for various regions and was compared with other researchers work in terms of precision and recall.

Jafarzadegan et al. [42] proposed the PCA-based combined hierarchical cluster approach and used the cophenetic correlation score and Wilcoxon hypothesis measure to compare the quality and efficiency of the cluster approach. Various datasets have been taken and compared with other existing methods and shown improvement using their methodology. Mateen et al. had proposed VGG-19-based feature extraction along with dimensionality reduction using PCA, SVD and optimized through Gaussian Mixture Model for the Fundus image classification and carried out the experimentation on the Kaggle image dataset consisting of 35,126 images and achieved more than 92% accuracy at various levels [43]. The authors of [44] presented keyword-based image recommendations for e-commerce products using the Markov-chain-based method along with the annotations for the image and results are compared with the existing methods. D. sha et al. had proposed a fashion clothing feature extraction for analyzing various attributes, such as cloth pattern, sleeves, and collar using GIST, Fourier and Pyramid Histogram of Oriented Gradients (PHOG) feature extraction and experimented using the tmall fashion dataset with 8000 images [45]. In [46], the authors proposed a text-based image recommendation using bag-of-words and the Term frequency–Inverse document frequency (TF-IDF) approach to get similar product image recommendations. For this approach, the authors experimented on the amazon fashion image dataset and computed the Euclidean distance measure to fetch the top-N similar recommendations.

Many researchers have worked on various machine learning approaches for similar image retrieval systems for various domains with supervised and unsupervised approaches. The majority of the researchers performed computing performance metrics such as precision, recall, and F1, where ground truth labels are known. In this paper, we have addressed a similar image recommender system for the e-commerce domain using various machine learning unsupervised clustering algorithms and computed the cluster performance metrics where ground truth labels are unknown.

3. Similar Image Recommendation

In this section, we discuss our proposed research method using Machine learning statistical unsupervised dimensionality reduction techniques using Principal Component Analysis through singular value decomposition (PCA-SVD) on a fashion image dataset and further applied unsupervised clustering algorithm using K-means(++) to find out the similar image clusters and computed the distance measures for the similar top-N product image recommendations and in Figure 2 shown our proposed research method.

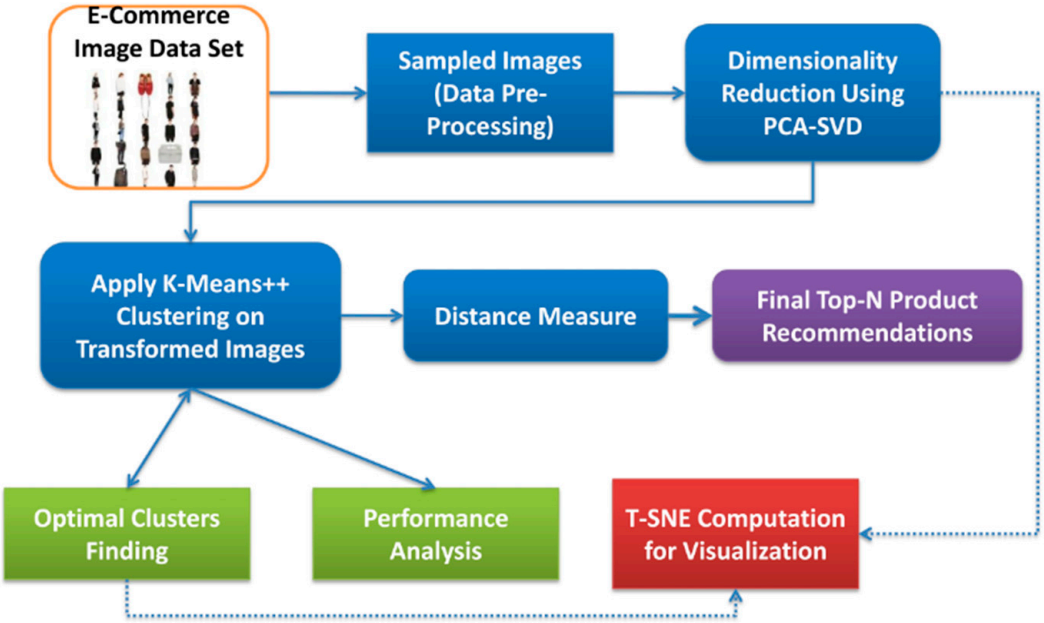

Figure 2. Proposed visual recommender system using machine learning approaches.

3.1. PCA through Eigen Value Decomposition (PEVD)

Dimensionality reduction is helpful to reduce the image dimensions and initial image matrix represented with the RGB (Red, Blue, and Green) color space values. Initially, the standardization process was for scaling and centering the data on a d-dimensional apparel image dataset. In this process, we can compute mean and standard deviation on the given data points as follows Equations (5)–(7) [9,47].

$$S_{mean} = Mean(s_i)_{i=1^n} \tag{5}$$

$$\sigma = Sd\_dev(s_i)_{i=1^n} \tag{6}$$

$$s_i^{'} = \frac{s_i - s_{mean}}{\sigma} \tag{7}$$

where s_i are the feature values of the given data matrix; further, the covariance data matrix is constructed to see how far the features are changing together by using the following covariance Equation (8), where A, B are the features and a_i, b_i are the feature values, and the mean is represented by μ.

$$cov(A, B) = \frac{1}{n}\sum_{i=1}^{n}(a_i - \mu_A) * (b_i - \mu_Y) \tag{8}$$

Further, the covariance matrix is used for calculating and building eigenvalues and eigenvectors for measuring the spread of the variance on projected vector points. Computing the eigenvalues and eigenvector from the following Equation (9):

$$\lambda_n v_n = U v_n \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_nandv_1, v_2, v_3, \ldots v_n$ are the eigenvalues and eigenvectors of U covariance matrix and further generates the new, transformed dimension vector.

3.2. PCA through Singular Value Decomposition (PSVD)

Principal components (PC) computation through singular value decomposition is a better approach in terms of numerical precision and stability. SVD [48,49] relies on the divide-and-conquer strategy, and eigenvalue decomposition mostly relies on the QR algorithm [50], which is less stable, and a loss

of precision can cause the forming of the covariance matrix in the EVD approach. PCA through SVD is also computationally efficient when compared to EVD in terms of dealing with the high dimensional datasets. SVD for principal components can be computed using Equation (10), where the given Matrix X singular values $\left(\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \sigma_4 \ldots . \sigma_p \geq 0\right)$ are represented by the Σ, and orthogonal values are represented by the $Uand\ V$, respectively.

$$X = U_{n\times m}\Sigma_{n\times p}V^T_{p\times p} \tag{10}$$

Furthermore, in our approach, we have performed the dimensionality reduction on our e-commerce dataset. The total dimensions in our dataset before PSVD [51] transformation is 14,400 and after applying the SVD and reduced by 90.01% and retained the 144 components to transform the e-commerce images. The following Figure 3 shows the reconstructed images from the original image after retaining the major principal components, and Figure 4 shown the 2-Dimensional visualization image projection for major principal components of the e-commerce product dataset.

Figure 3. Sample Principal Component Analysis (PCA) through Singular Value Decomposition (PSVD) transformed reconstructed images.

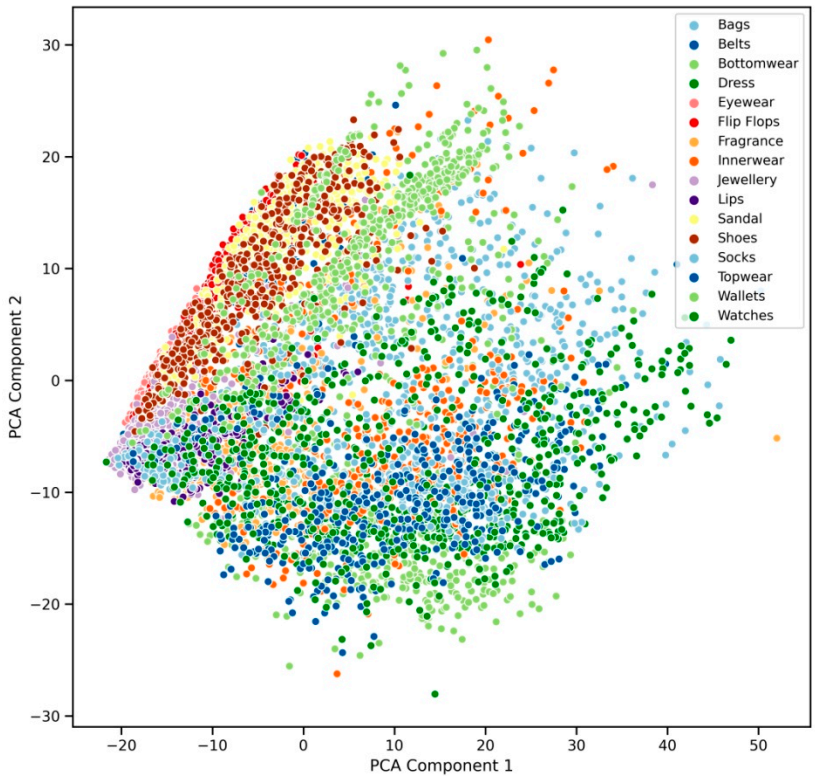

Figure 4. PCA-SVD Transformed images using 2-D visualization.

3.3. K-Mean(++) Clustering Approach

Clustering is an unsupervised learning technique to find out the K-Patterns in the given image dataset. For our image similarity approach, we have taken a Fashion image dataset with 40,000 images for analysis. We have computed K-Mean clustering [52] by computing ten iterative times with different centroid positions, and the K-Means++ [53] initialization method has been used for better convergence. An earlier version of K-Means consists of the initialization method, assignment of data points, updating the centroid, and repeats the stages until it finds out the convergence. Hence, choosing the random k-centroids is also known as the Initialization sensitivity procedure. To overcome the initialization sensitivity procedure in our approach, we have adapted the K-Means ++ approach for finding out the better initial selection of K-centroids for PCA transformed images using the SVD approach on our e-commerce 40K dataset. Here, in the K-Means ++ approach initially picks a centroid point C_1 arbitrarily and computes the distance to all available data points from the arbitrarily initialized centroid, as shown in Equation (11).

$$d_i = max_{(j:1-m)} \left\| x_i - C_j \right\|^2 \tag{11}$$

where m is the number of centroids used and picked for each iteration and x_i becomes the new centroid depends on the proportion of higher probability to distance d_i and we repeat these steps until finding the K-centroids. Further, we find the Euclidean distance for each data point from the K-centroids using Equation (12).

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \tag{12}$$

Closest points with respect to the e-commerce images are assigned to the clusters and repeats this process until we reach the convergence and all the image points (S_i) are assigned to ith Cluster by using Equation (13).

$$C_{i=}\frac{1}{|S_i|}\sum_{x_i \in S_i} x_i \tag{13}$$

Further, we have generated 16 different clusters from sampled images from the 40K image dataset by computing inertia, as shown in Figure 5, and the silhouette coefficient to understand the goodness of our cluster model. Equation (14) is used to measure the goodness metric through the silhouette coefficient [54] on various clusters formed using k-Means++.

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} if a(i) < b(i) \\ 0 \quad if a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 \, if a(i) > b(i) \end{cases} \tag{14}$$

where $a(i)$, $b(i)$ are the average and lowest average distances between i and all other data points within and different clusters, respectively.

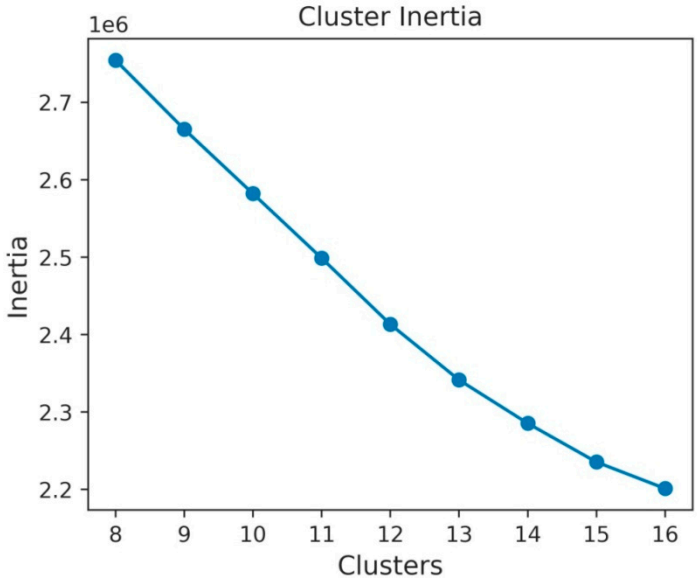

Figure 5. Optimal clusters finding using K-means++ inertia.

3.4. Similar Distance Measure

Fetching the relevant similar images from the given input image from the fashion dataset, we have computed the Manhattan distance measure [15] between any two given vectors X and Y to compute the lowest distance on PCA through SVD transformed image clusters generated using the K-Means++ approach for e-commerce product image clusters by using the distance metric function as shown in Equation (15) and fetches the top-5 most similar images for the given input image as user recommendations.

$$Manhattan(X, Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{15}$$

4. Dataset and Experimentation

We have collected 40K fashion product image dataset [55] along with the major category and subcategory of the products. There is a total of six major categories of products, such as apparel, accessories, footwear, personal care, sporting goods, etc., and 44 subcategories of the various fashion products top wear, shoes, bags, watches, wallets, belts, etc. Out of these major categories and subcategories, most items are listed in apparel, accessories, footwear, and personal care. Further, as in

the pre-processing data stage, we have sampled out into our training dataset with 7632 images of 16 subcategories for our experimentation purpose with each category of equal distribution to avoid the unbalance data distribution problems that affect the final recommendations. Figure 6 shows the category product image distribution from our fashion product image dataset.

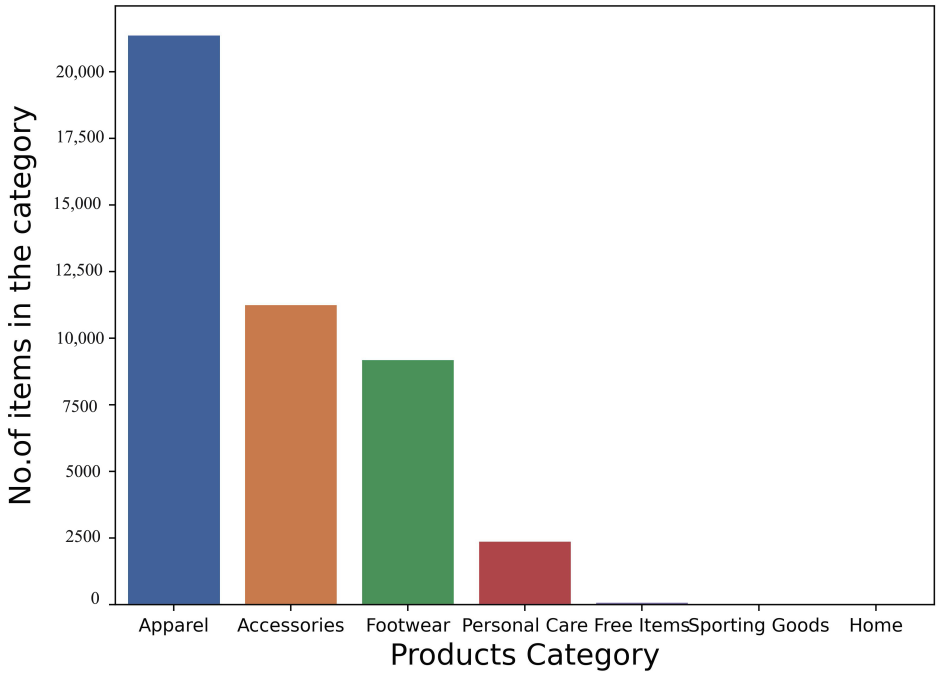

Figure 6. Total category distribution from the Kaggle fashion product dataset.

Experimentation

In this section, we performed our experimentation using the PSVD approach for dimensionality reduction and fetched the important principal components. Initially, we have set the number of components to 500 and computed the PCA singular values through the PSVD process on our sampled dataset dimensionally reduced to 90.01% generates 144 final principal components. Figure 7 shows cumulative variance by 500 PCA components and convergence achieved at 144 principal components.

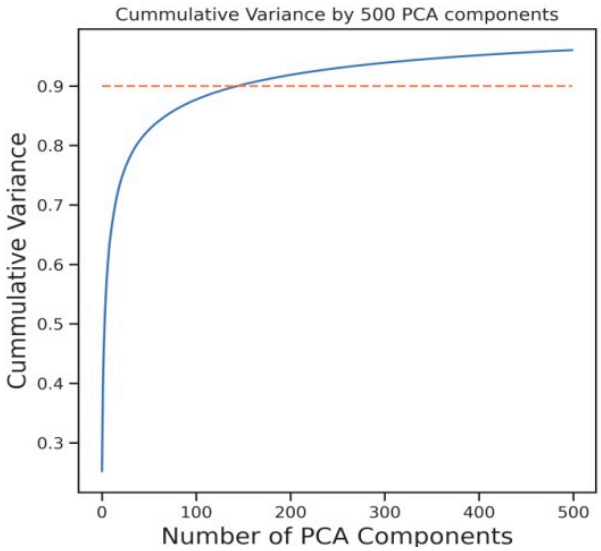

Figure 7. PCA through SVD cumulative variance.

Further, we have computed the K-means++ clustering approach on the PSVD transformed images. Here, we have computed various initializations for the number of cluster identifications ranging from 8 to 16. For each iteration, we tried to determine the better convergence for the better number of cluster formation of our fashion products. Further, we have calculated each cluster inertia to maximize the quality estimation of the number of clusters for further analysis. We have found optimal values at cluster 16 with a lower inertia value, which converge after 35 iterations. In Figures 8 and 9, showed the high dimensional cluster segregation for 16 subcategories of images using t-distributed stochastic neighbor embedding (t-SNE) [56] for the K-means++ approach.

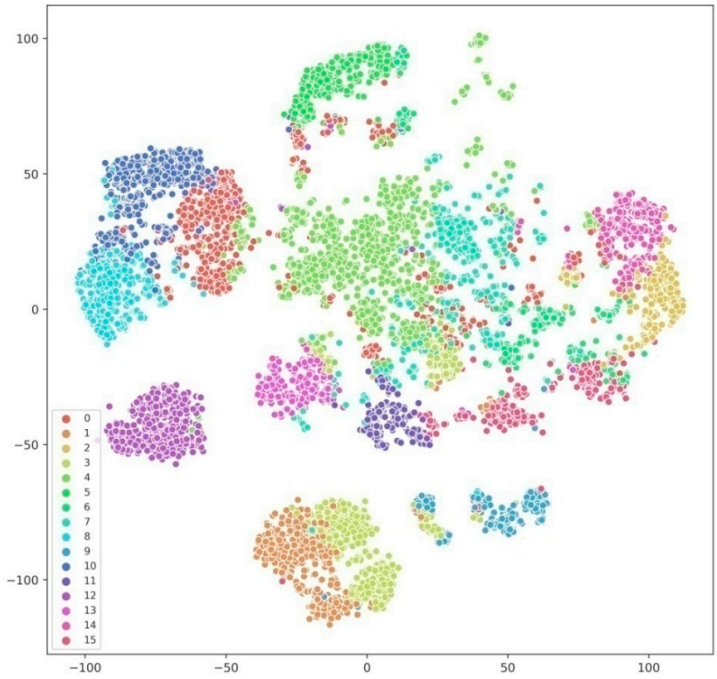

Figure 8. T-distributed stochastic neighbor embedding (T-SNE) visualization on PSVD transformed using K-means++.

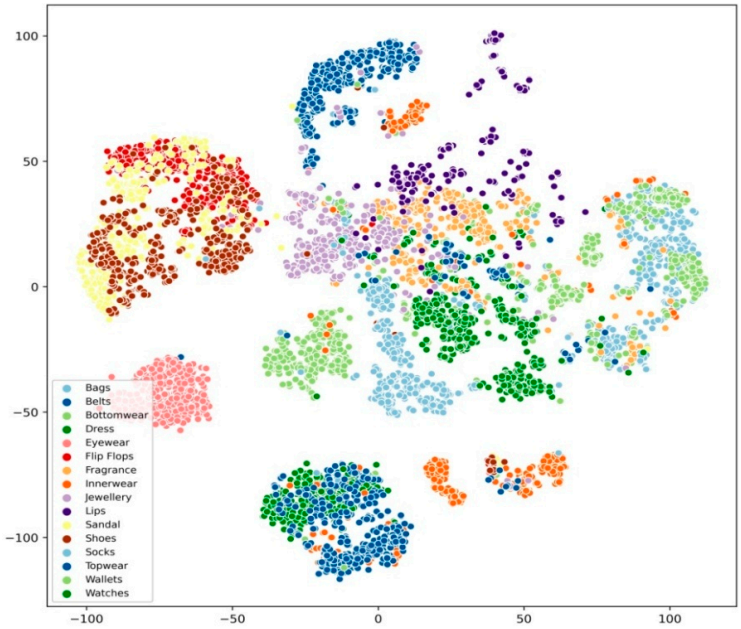

Figure 9. T-SNE visualization on the original dataset.

Finally, we computed the distance measure using the Manhattan distance measure for the given input image to the cluster images and fetched the top-K similar images, and we have further analyzed and compared with other standard clustering algorithms for better results. All our experimentations are carried out using a desktop computer with Intel Core i7-8700K Processor/16 GB of RAM/64-bit Ubuntu operating system/Python 3.6 Environment.

5. Results

We have carried out various cluster performance metrics to analyze and achieve better recommendation results for our approach. Apart from inertia for K-means clustering, we have, furthermore, computed various cluster performance metrics, such as the silhouette coefficient, as computed using Equation (14) with results ranges from the values −1 to 1, where the higher the SC value, the more efficient the cluster. In our approach, we have achieved a 0.1414 coefficient value, which is a higher value from the other algorithms. The comparison results for the SC coefficient for various unsupervised clustering algorithms are shown in Figure 10.

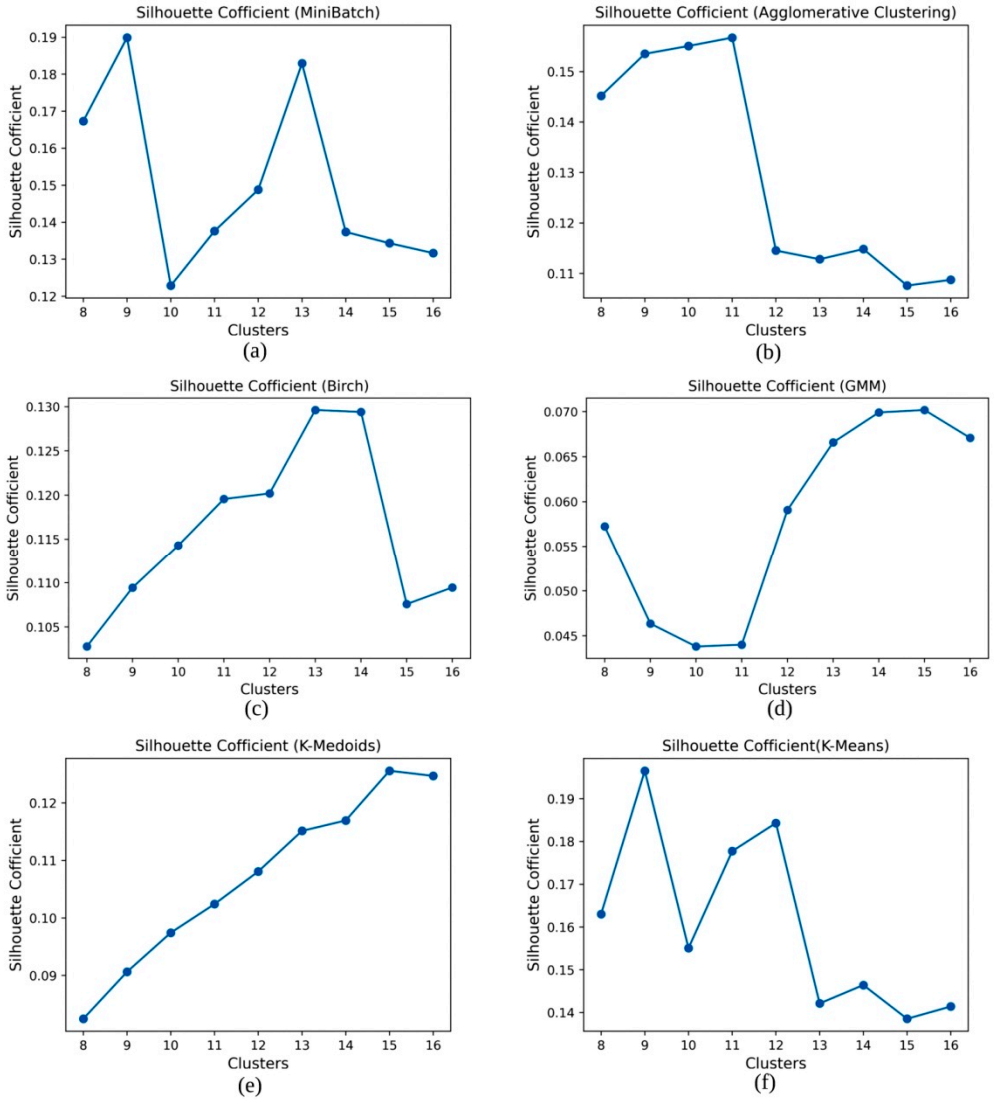

Figure 10. Comparison of the silhouette coefficient of various clusters shown in (**a**–**f**).

To analyze the criteria for the variance ratio for checking the clusters, we computed the Calinski-Harabasz (CH) score. The CH score tells us whether or not the clusters are well separated

and if the optimal value is higher than any other clustering approach. Equations (16)–(18) shows the calculation of the CH score:

$$CH(k) = \left[\frac{B(k)}{W(k)}\right] \times \left[\frac{(n-k)}{(k-1)}\right] \tag{16}$$

$$W(k) = \sum_{q=1}^{k} \sum_{x \in C_q} \left(x - c_q\right)\left(x - c_q\right)^2 \tag{17}$$

$$W(k) = \sum_{q=1}^{k} \sum_{x \in C_q} \left(c_q - c_E\right)\left(c_q - c_E\right)^2 \tag{18}$$

where n is the number of data points, $B(K)$, $W(K)$ are between and within the cluster variation, respectively, k is total clusters, c_q is the set of cluster points, and the total data size is E. The CH score we got here for our approach is 669.4, which an optimal value compared to other algorithms. The comparison results for various CH scores are shown in Figure 11.

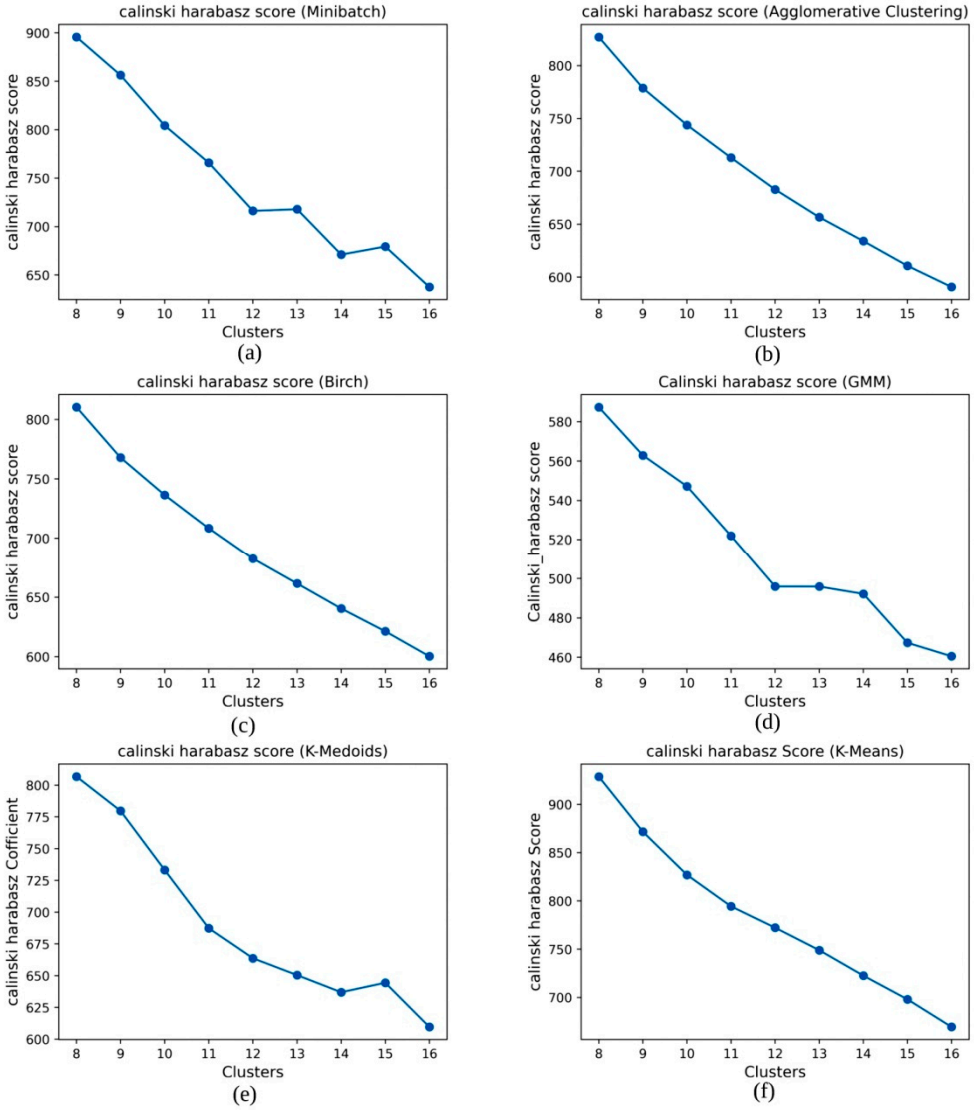

Figure 11. Comparison of the Calinski-Harabasz score of various clusters shown in (**a**–**f**).

Finally, we have computed the Davies–Bouldin index score for the average cluster similarity, which computes that an optimal value is a high score among the clusters, and zero is the lowest score. The DB index score can be computed using Equations (19) and (20):

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j} M_{ij} \tag{19}$$

$$M_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{20}$$

where s is the mean distance between cluster points and M is to compute the similarity between centroids of the clusters i and j. The DB score we achieved is 1.8538, which is slightly lower compared to other algorithms. The comparative performance of the DB index score for various clustering algorithms is shown in Figure 12.

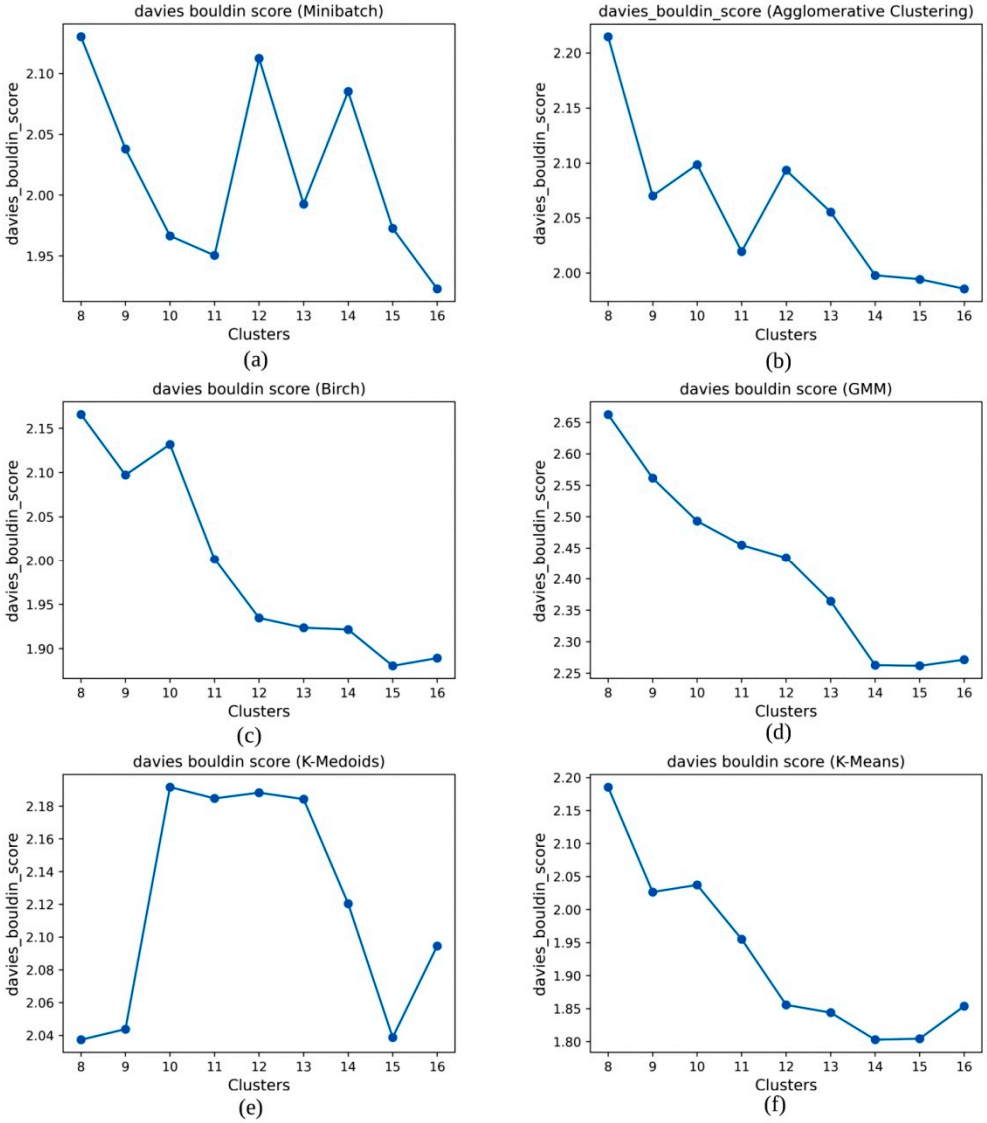

Figure 12. Comparison of Davies–Bouldin (DB) index score of various clusters shown in (**a**–**f**).

Tables 1 and 2 show the PSVD transformed cluster evaluations for various standard algorithms and computational time comparisons for various clustering algorithms.

Table 1. Performance comparison of PSVD transformed clustering algorithms.

Clustering Algorithm	SC Coefficient	CH Score	DB Score
MiniBatch	0.131669	637.60	1.9229
K-Mediod	0.124709	609.61	2.0946
Agglomerative	0.108739	590.66	1.9856
Brich	0.109466	600.36	1.8893
GMM	0.067083	460.55	2.2714
K-Means++	**0.141421**	**669.44**	**1.8538**

Table 2. Computational wall time of various clustering algorithms.

Clustering Algorithm	Computation Wall Time (Milliseconds (ms)/Seconds (s)/Minutes (min)			
	Cluster Fitting	SC Coefficient	CH Score	DB Score
MiniBatch	2.51 s	9.09 s	51.2 ms	120 ms
K-Mediod	10.9 s	9.08 s	60. 4 ms	108 ms
Agglomerative	40.7 s	11.6 s	79.1 ms	131 ms
Brich	35.4 s	9.05 s	48.2 ms	108 ms
GMM	1 min 4 s	10.2 s	1.16 s	1.12 s
K-Means++	21.3 s	48.5 s	52.5 ms	1.04 s

Figure 13 shown the final similar product recommendations using PSVD transformed K-means++ approach with Top-5 recommendations fetched using the Manhattan distance measure.

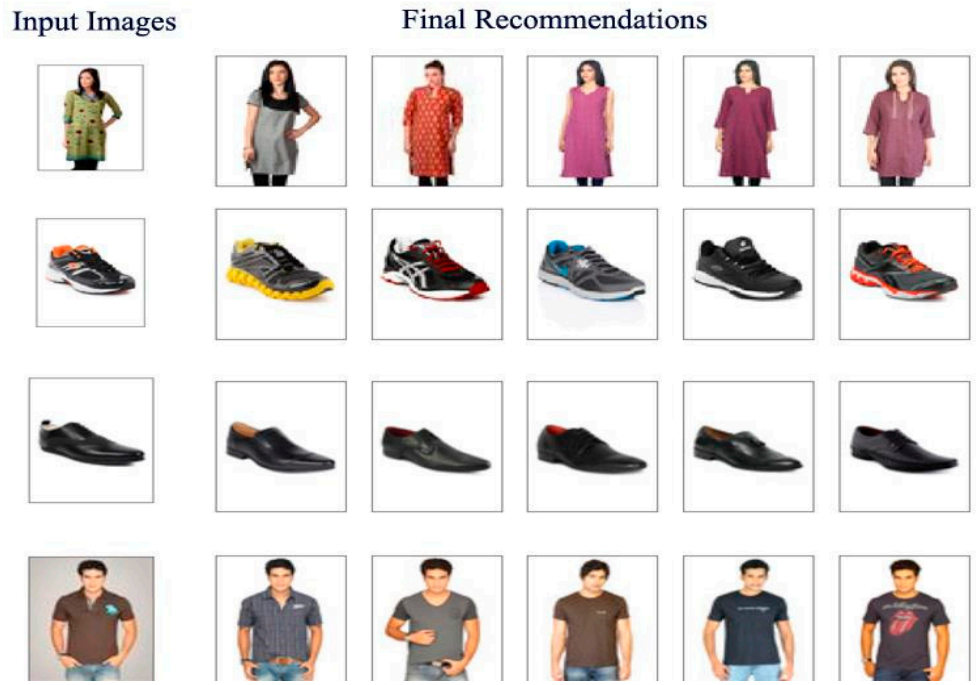

Figure 13. Top-5 final recommendations from the proposed approach.

6. Conclusions

The purpose of our proposed model is to fetch similar images for e-commerce portals when the user selects the desired product image using unsupervised statistical machine learning techniques. Here, in our approach, we initially performed PSVD dimensionality reduction on the fashion product image dataset and retained the 144 principal components and achieved a 90.01% variance from the

original dimensions, which are a total of 14,400 dimensions. Further, we performed the K-means++ clustering on the PSVD transformed images. We have compared PSVD transformed with other clustering algorithms, such as MiniBatch, Agglomerative, Brich, the Gaussian Mixture Model (GMM), and K-Mediod. Performance metrics are evaluated on all these algorithms for the SC coefficient for similarity, CH score for variance ration, and DB score for average similarity. Out of all the measures, PSVD-K-means++ has scored well on the SC coefficient and CH score and is lagging in the DB index score. Where we observed in the final similar image recommendations, most of the recommendations work well, but, in a case like if the input image orientation changes, it is possible to get the mixed product suggestion instead of all the similar images. To overcome these limitations further, we suggested applying image augmentation techniques and deep learning approaches for more accurate image feature extraction and training for similar image recommendations.

Author Contributions: Conceptualization, methodology, validation, writing—original draft preparation, S.K.A.; Supervision, review and guidance, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wulansaria, I.Y.; Parwantob, N.B. Asian E-Commerce Engages Global Trade Openness: The Role of Information and Communications Technology, Social, and Security Indicators. *Int. J. Innov. Creat. Chang.* **2020**, *11*, 12.
2. eCommerce—Asia | Statista Market Forecast. Available online: https://www.statista.com/outlook/243/101/ecommerce/asia (accessed on 2 February 2020).
3. Salau, A.O.; Jain, S. Feature Extraction: A Survey of the Types, Techniques, Applications. In Proceedings of the 2019 International Conference on Signal Processing and Communication (ICSC), Noida, India, 7–9 March 2019; pp. 158–164.
4. Haji, M.S.; Alkawaz, M.H.; Rehman, A.; Saba, T. Content-Based Image Retrieval: A Deep Look at Features Prospectus. *Int. J. Comput. Vis. Robot.* **2019**, *9*, 14–38. [CrossRef]
5. Kumari, M. Content Based Image Retrieval. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3371777 (accessed on 14 May 2019).
6. Zhou, J.; Liu, X.; Liu, W.; Gan, J. Image Retrieval Based on Effective Feature Extraction and Diffusion Process. *Multimed. Tools Appl.* **2019**, *78*, 6163–6190. [CrossRef]
7. Pal, M.S.; Garg, D.S.K. Image Retrieval: A Literature Review. *Int. J. Adv. Res. Comput. Eng. Technol.* **2013**, *2*, 1323–2278.
8. Limitations of Text Based Image Retrieval Psychology Essay. Available online: https://www.ukessays.com/essays/psychology/limitations-of-text-based-image-retrieval-psychology-essay.php (accessed on 15 October 2020).
9. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
10. Roweis, S.T. EM Algorithms for PCA and SPCA. *Advances in Neural Information Processing Systems*. pp. 626–632. Available online: http://papers.neurips.cc/paper/1398-em-algorithms-for-pca-and-spca (accessed on 22 October 2020).
11. Tharwat, A.; Gaber, T.; Ibrahim, A.; Hassanien, A.E. Linear Discriminant Analysis: A Detailed Tutorial. *AI Commun.* **2017**, *30*, 169–190. [CrossRef]
12. Tipping, M.E.; Bishop, C.M. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [CrossRef]
13. Kaya, I.E.; Pehlivanlı, A.Ç.; Sekizkardeş, E.G.; Ibrikci, T. PCA Based Clustering for Brain Tumor Segmentation of T1w MRI Images. *Comput. Methods Programs Biomed.* **2017**, *140*, 19–28. [CrossRef]
14. Geng, X.; Zhang, H.; Bian, J.; Chua, T.-S. Learning Image and User Features for Recommendation in Social Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7 December 2015; pp. 4274–4282.

15. Tyagi, V. Similarity Measures and Performance Evaluation. In *Content-Based Image Retrieval*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 63–83.
16. Clustering Scikit-Learn 0.23.2 documentation. Available online: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation (accessed on 2 February 2020).
17. Steinley, D.; Brusco, M.J.; Hubert, L. Properties of the Hubert-Arable Adjusted Rand Index. *Psychol. Methods* **2004**, *21*, 261. [CrossRef]
18. Rosenberg, A.; Hirschberg, J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007; pp. 410–420.
19. Hirschberg, J.B.; Rosenberg, A. V-Measure: A Conditional Entropy-Based External Cluster Evaluation. 2007. Available online: http://www.aclweb.org/anthologyy/D07-103 (accessed on 14 August 2020).
20. Gholamian, M.; Jahanpour, S.; Sadatrasoul, S. A New Method for Clustering in Credit Scoring Problems. *J. Math. Comput. Sci.* **2013**, *6*, 97–106. [CrossRef]
21. Lu, Y.; Wu, Y.; Liu, J.; Li, J.; Zhang, P. Understanding Health Care Social Media Use from Different Stakeholder Perspectives: A Content Analysis of an Online Health Community. *J. Med. Internet Res.* **2017**, *19*, e109. [CrossRef] [PubMed]
22. Hancer, E.; Ozturk, C.; Karaboga, D. Artificial Bee Colony Based Image Clustering Method. In Proceedings of the 2012 IEEE Congress on Evolutionary Computation, Brisbane, Australia, 10–15 June 2012; pp. 1–5. [CrossRef]
23. Gong, M.; Liang, Y.; Shi, J.; Ma, W.; Ma, J. Fuzzy C-Means Clustering with Local Information and Kernel Metric for Image Segmentation. *IEEE Trans. Image Process.* **2012**, *22*, 573–584. [CrossRef] [PubMed]
24. Karthikeyan, M.; Aruna, P. Probability Based Document Clustering and Image Clustering Using Content-Based Image Retrieval. *Appl. Soft Comput.* **2013**, *13*, 959–966. [CrossRef]
25. Zeng, K.; Yu, J.; Li, C.; You, J.; Jin, T. Image Clustering by Hyper-Graph Regularized Non-Negative Matrix Factorization. *Neurocomputing* **2014**, *138*, 209–217. [CrossRef]
26. Ozturk, C.; Hancer, E.; Karaboga, D. Improved Clustering Criterion for Image Clustering with Artificial Bee Colony Algorithm. *Pattern Anal. Appl.* **2015**, *18*, 587–599. [CrossRef]
27. Younus, Z.S.; Mohamad, D.; Saba, T.; Alkawaz, M.H.; Rehman, A.; Al-Rodhaan, M.; Al-Dhelaan, A. Content-Based Image Retrieval Using PSO and k-Means Clustering Algorithm. *Arab. J. Geosci.* **2015**, *8*, 6211–6224. [CrossRef]
28. Lin, X.; Li, C.-T. Large-Scale Image Clustering Based on Camera Fingerprints. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 793–808. [CrossRef]
29. Villalba, L.J.G.; Orozco, A.L.S.; Corripio, J.R. Smartphone Image Clustering. *Expert Syst. Appl.* **2015**, *42*, 1927–1940. [CrossRef]
30. Yang, J.; Parikh, D.; Batra, D. Joint Unsupervised Learning of Deep Representations and Image Clusters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
31. Pandey, S.; Khanna, P. Content-Based Image Retrieval Embedded with Agglomerative Clustering Built on Information Loss. *Comput. Electr. Eng.* **2016**, *54*, 506–521. [CrossRef]
32. L i, S.; Yu, H.; Yuan, L. A Novel Approach to Remote Sensing Image Retrieval with Multi-Feature vp-Tree Indexing and Online Feature Selection. In Proceedings of the 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), Taipei, Taiwan, 20–22 April 2016; pp. 133–136. [CrossRef]
33. Biradar, M.; Ahmed, M. Visual Based Information Retrieval Using Voronoi Tree. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, Beijing, China, 2–5 November 2017; pp. 603–609.
34. Zhu, X.; Li, X.; Zhang, S.; Xu, Z.; Yu, L.; Wang, C. Graph PCA Hashing for Similarity Search. *IEEE Trans. Multimed.* **2017**, *19*, 2033–2044. [CrossRef]
35. Fachrurrozi, M.; Fiqih, A.; Saputra, B.R.; Algani, R.; Primanita, A. Content Based Image Retrieval for Multi-Objects Fruits Recognition Using k-Means and k-Nearest Neighbor. In Proceedings of the 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, Indonesia, 1–2 November 2017; pp. 1–6.

36. Yildiz, K. Dimensionality Reduction-Based Feature Extraction and Classification on Fleece Fabric Images. *Signal Image Video Process.* **2017**, *11*, 317–323. [CrossRef]
37. Cui, Y.; Wong, W.K. Textile Image Retrieval Using Joint Local PCA-Based Feature Descriptor. In *Applications of Computer Vision in Fashion and Textiles*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 253–271.
38. Chen, Y.; Tang, S.; Bouguila, N.; Wang, C.; Du, J.; Li, H.L. A Fast Clustering Algorithm Based on Pruning Unnecessary Distance Computations in DBSCAN for High-Dimensional Data. *Pattern Recognit.* **2018**, *83*, 375–387. [CrossRef]
39. Singh, V.P.; Srivastava, R. Improved Image Retrieval Using Fast Colour-Texture Features with Varying Weighted Similarity Measure and Random Forests. *Multimed. Tools Appl.* **2018**, *77*, 14435–14460. [CrossRef]
40. Ashraf, R.; Ahmed, M.; Jabbar, S.; Khalid, S.; Ahmad, A.; Din, S.; Jeon, G. Content Based Image Retrieval by Using Color Descriptor and Discrete Wavelet Transform. *J. Med. Syst.* **2018**, *42*, 44. [CrossRef] [PubMed]
41. Jian, M.; Yin, Y.; Dong, J.; Lam, K.-M. Content-Based Image Retrieval via a Hierarchical-Local-Feature Extraction Scheme. *Multimed. Tools Appl.* **2018**, *77*, 29099–29117. [CrossRef]
42. Jafarzadegan, M.; Safi-Esfahani, F.; Beheshti, Z. Combining Hierarchical Clustering Approaches Using the PCA Method. *Expert Syst. Appl.* **2019**, *137*, 1–10. [CrossRef]
43. Mateen, M.; Wen, J.; Nasrullah; Song, S.; Huang, Z. Fundus Image Classification Using VGG-19 Architecture with PCA and SVD. *Symmetry* **2019**, *11*, 1. [CrossRef]
44. Sejal, D.; Rashmi, V.; Venugopal, K.R.; Iyengar, S.S.; Patnaik, L.M. Image Recommendation Based on Keyword Relevance Using Absorbing Markov Chain and Image Features. *Int. J. Multimed. Inf. Retr.* **2016**, *5*, 185–199. [CrossRef]
45. Sha, D.; Wang, D.; Zhou, X.; Feng, S.; Zhang, Y.; Yu, G. An Approach for Clothing Recommendation Based on Multiple Image Attributes. In Proceedings of the International Conference on Web-Age Information Management, Nanchang, China, 3–5 June 2016; pp. 272–285.
46. Shrivastava, R.; Sisodia, D.S. Product Recommendations Using Textual Similarity Based Learning Models. In Proceedings of the 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 23–25 January 2019; pp. 1–7.
47. Aït-Sahalia, Y.; Xiu, D. Principal Component Analysis of High-Frequency Data. *J. Am. Stat. Assoc.* **2019**, *114*, 287–303. [CrossRef]
48. Furnas, G.W.; Deerwester, S.; Durnais, S.T.; Landauer, T.K.; Harshman, R.A.; Streeter, L.A.; Lochbaum, K.E. Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure. *ACM SIGIR Forum* **2017**, *51*, 90–105. [CrossRef]
49. De Lathauwer, L.; De Moor, B.; Vandewalle, J. A Multilinear Singular Value Decomposition. *SIAM J. Matrix Anal. Appl.* **2000**, *21*, 1253–1278. [CrossRef]
50. Parlett, B.N. The QR Algorithm. *Comput. Sci. Eng.* **2000**, *2*, 38–42. [CrossRef]
51. Wall, M.E.; Rechtsteiner, A.; Rocha, L.M. Singular Value Decomposition and Principal Component Analysis. In *A Practical Approach to Microarray Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 91–109.
52. Ding, C.; He, X. K-Means Clustering via Principal Component Analysis. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 29.
53. Arthur, D.; Vassilvitskii, S. *K-Means++: The Advantages of Careful Seeding*; Stanford University: Stanford, CA, USA, 2006; pp. 1–11.
54. Aranganayagi, S.; Thangavel, K. Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure. In Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007), Sivakasi, India, 13–15 December 2007; Volume 2, pp. 13–17.
55. Fashion Product Images Dataset | Kaggle. Available online: https://www.kaggle.com/paramaggarwal/fashion-product-images-dataset (accessed on 2 February 2020).
56. van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.