Erzincan Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
2024, 17(3), 638-649
ISSN: 1307-9085, e-ISSN: 2149-4584
Araştırma Makalesi

Erzincan University
Journal of Science and Technology
2024, 17(3), 638-649
DOI: 10.18185/erzifbed.1500279
Research Article

Efficient Image Retrieval in Fashion: Leveraging Clustering and Principal Component Analysis for Search Space Reduction

Başak Esin Köktürk-Güzel[1,2*]

[1]İzmir Democracy University, Electrical and Electronics Engineering, Gürsel Aksel Bulvarı No:14
Karabağlar/İZMİR 35350
[2]Zoi Data Yazılım Danışmanlık Ltd.Şti. İzmir Teknoloji Geliştirme Bölgesi A8 Binası No:8
Urla / İZMİR 34430

Received: 13/06/2024, **Revised:** 05/09/2024, **Accepted:** 16/09/2024, **Published:** 31/12/2024

Abstract

In this study, a novel approach using clustering techniques and Principal Component Analysis (PCA) for reducing the search space in fashion image retrieval systems is introduced. The study focuses on extracting high-dimensional feature vectors from images of clothing items and finding the same or the most similar product using these feature vectors by narrowing the search space. The proposed method employs unsupervised learning algorithms to analyze high-dimensional fashion image feature vectors, grouping them into meaningful clusters. This enhances search efficiency and improves user experience. By reducing the dimensionality of feature vectors with PCA, computational costs are reduced. Experimental results demonstrate that the proposed method significantly improves computation time while maintaining an acceptable level of accuracy.

Keywords: Fashion search, fashion image retrieval, clustering, principal component analysis, Resnet50, VGG19

Modada Etkili Görüntü Arama: Arama Alanını Azaltmak İçin Kümeleme ve Temel Bileşen Analizinden Yararlanma

Öz

Bu çalışmada, moda görüntü arama sistemlerinde arama alanını küçültmek için kümeleme teknikleri ve Temel Bileşen Analizi (PCA) kullanan yeni bir yaklaşım tanıtılmıştır. Çalışmada kıyafet ürünlerine ait görüntülerden yüksek boyutlu öznitelik vektörleri çıkarılmış ve bu öznitelik vektörleri kullanılarak aynı ya da en benzer ürünün bulunması ve arama uzayının daraltılması ele alınmıştır. Önerilen yöntem yüksek boyutlu moda görüntü özellik vektörlerini analiz etmek için denetimsiz öğrenme algoritmaları kullanarak, onları anlamlı kümelere ayırmaktadır. Bu sayede arama verimliliği artırılmakta ve kullanıcı deneyimi iyileştirilmektedir. PCA ile öznitelik vektör boyutları indirgenerek hesaplama maliyeti azaltılmıştır. Deneysel sonuçlar, önerilen yöntemin hesaplama süresini önemli ölçüde düşürürken kabul edilebilir bir doğruluk seviyesini koruduğunu göstermektedir.

Anahtar Kelimeler: Moda arama, moda görüntü arama, kümeleme, temel bileşen analizi, Resnet50, VGG19

*Corresponding Author: basak.guzel@idu.edu.tr
Başak Esin KÖKTÜRK GÜZEL, https://orcid.org/ 0000-0002-9429-1149

1. Introduction

Finding a product in the fashion world can become a challenging problem with the growing amount of data as e-commerce sites upload thousands of new product images every day along with their textual descriptions. The increasing diversity and production speed in the fashion industry make it challenging for customers to choose the right product. In response to these challenges, the use of AI techniques to ensure visual feature compatibility and provide accurate product recommendations is of great importance in the fashion domain [1]. Traditionally, users on e-commerce sites search for products by entering relevant keywords into a search box. The search algorithm matches these keywords with product tags in the database and presents the relevant products to the user. For text-based searches to be effective, it's important for the customer to fully understand/identify the product and know the appropriate keywords to enter. There can be many possible descriptions for an image, making it difficult for a search engine to precisely identify a product, and adding a close-up of the product is much easier. People are good at recognizing images and tend to think visually. Using visual search, customers can instead search for a product using images [2], [3]. Visual search allows customers to find information faster using images instead of words.

A fashion image retrieval task is to find query images or similar images in a database. Despite advancements in technology and deep learning methods, there are specific limitations in real-world applications. A few of these limitations include taking images under uncontrolled circumstances (different lighting, different angles, etc.), displaying multiple fashion products in a single image with some not fully visible (for instance, only half of the model's trousers are visible in a t-shirt image) and being sensitive to shape deformations. [4]. Therefore, there are still problems that need to be solved and automated specifically for fashion images.

In the literature, various studies exist on segmenting and classifying fashion products through images [5], [6], [7] or videos [8], [9]. The methods proposed in these studies have been trained with large, labeled datasets and have utilized different classifiers to create tags for test data. Today, most e-commerce sites still aim to find similar products by creating tags based on images and searching their databases. However, approaching the problem of finding similar products based on images in this way generally requires defining the products with fixed classifiers such as category, collar type, sleeve length, material, length, etc. Yet, searching through these classifiers to find similar products may not always provide the result the user is looking for. Typically, when someone performs a visual search for a product online, they want to find and purchase that exact product. Therefore, searching across images becomes important. For this purpose, Huang et.al. proposed a Dual-attribute perceptual Ranking Network (DARN) for feature learning based on the Siamese network [10]. Also, Berg et.al. used CNN network to find similar product in shop where the input query is a street photo [11]. However, these methods are early research efforts in this innovative area and require further refinement to achieve more precise alignment and improve computational efficiency.

In addition to finding the most similar products, the computational cost associated with image searches can grow exponentially as the number of images in the database increases. To tackle this issue, our study suggests reducing the search space through image clustering. We further investigate the potential of feature reduction via Principal Component Analysis (PCA) to decrease the costs associated with calculating distances between images. This method not only makes the search process more streamlined but also improves efficiency, enabling the effective management of larger datasets.

2. Materials and Methods

In this study, initially, the input image undergoes segmentation to isolate the clothing item from the background. This segmented image is then processed through a deep convolutional neural network, either ResNet50 or VGG19, to extract high-dimensional feature vectors representing the visual characteristics of the item. To manage the high dimensionality and improve computational efficiency, Principal Component Analysis (PCA) is applied to reduce the feature vector dimensions. Subsequently, the reduced feature vector can be optionally assigned to a pre-defined cluster, which helps in organizing the feature space and potentially speeds up the similarity search. This clustering step is particularly useful for large-scale databases, where it can significantly enhance retrieval performance by narrowing down the candidate items. Finally, cosine similarity is calculated between the reduced input feature vector and the reduced feature vectors of candidate items in the relevant clusters to identify the most visually similar items. This approach ensures efficient and accurate retrieval of similar clothing images, facilitating enhanced user experience in fashion-related applications. The block diagram of the proposed method is illustrated in Figure 1, and each block is explained in detail in the subsequent sections.

2.1.Dataset

In this study, BeautifulSoup (BS4) Python library [12] has been used to scrape web pages. Bs4 allows us to read and manipulate HTML codes, as well as search within tags in HTML code. Since each website has its own HTML structure, searching via tags facilitates easier access to images and essential information.

For this project, HTML codes from four major e-commerce websites operating in the global market were scraped using the Bs4 library to extract images of clothing items and related information. A variety of details were extracted from the data, including the title of the product, a hyperlink to the product's image, categories, features, and price. For this research we have only used product images. We have collected 41303 images and since each image has more than one fashion item we have extracted total 81702 product images by segmentation of the fashion products.

All images were downloaded to a local machine to prevent the links from becoming inactive if the products were removed from the website.

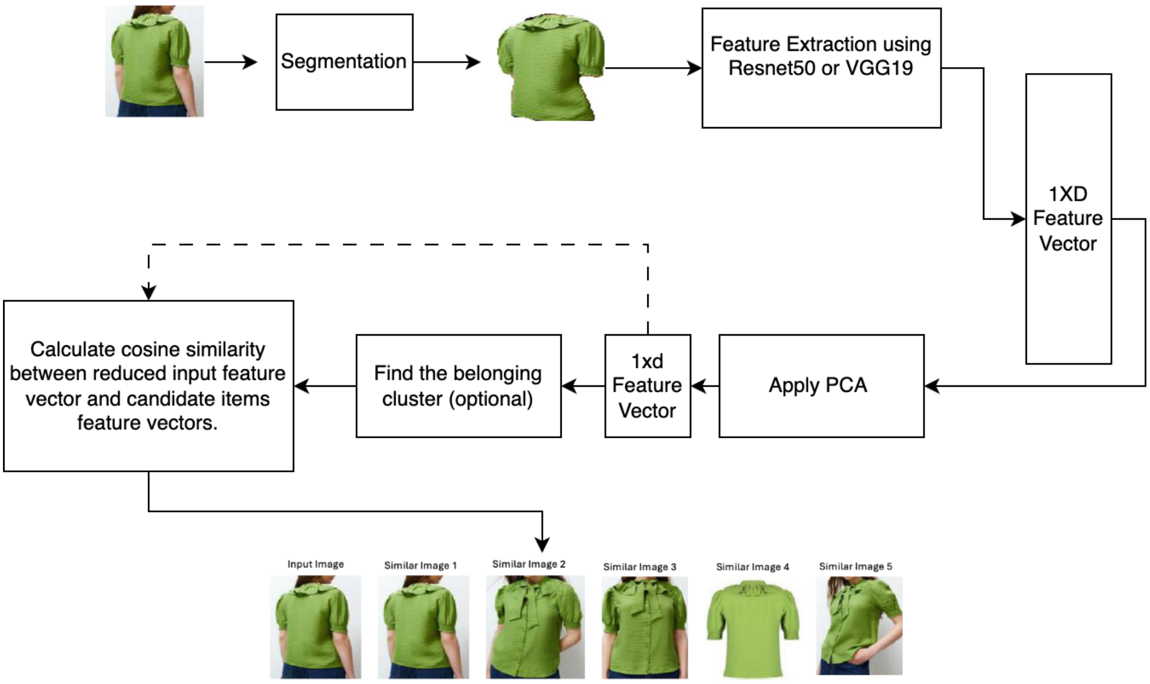

Segmentation
Feature Extraction using Resnet50 or VGG19
1XD Feature Vector
Apply PCA
1xd Feature Vector
Find the belonging cluster (optional)
Calculate cosine similarity between reduced input feature vector and candidate items feature vectors.
Input Image
Similar Image 1
Similar Image 2
Similar Image 3
Similar Image 4
Similar Image 5

Figure 1Block Diagram of the Proposed Method

2.2.Segmentation of Fashion Items

Image segmentation is the process of separating different features in an image into distinct regions so they can be differentiated. In this study, we used image segmentation methods to identify a fashion product. Detectron2 [13] is a computer vision tool that can spot objects, segment images, and estimate poses using extensible tools from Facebook AI Research (FAIR). Pre-trained models in Detectron2 save time and compute resources; a modular design simplifies model customization; and its scalable architecture guarantees high performance on large datasets. We tested different image segmentation models like U-net, Attention-Unet, and MaskR-CNN. Detectron2 beats out all the others after comparison evaluations.

Image segmentation methods generally evaluated by calculating Intersection-over-Union (IoU). IoU quantifies the extent of overlap between the predicted region Y and the true region X. The IoU metric returns a value between 0 and 1, where 0 indicates no overlap and 1 indicates perfect segmentation. IoU is given by:

$$IoU = \frac{|X \cap Y|}{|X \cup Y|}$$

We trained and tested various segmentation models on the iMaterialist dataset, which was originally prepared for an image processing competition at CVPR2019 [14]. The dataset contains 6,760 images: 5,408 for training, 676 for testing, and 676 for validation. To ensure an equal distribution of the reduced categories, we used a subset of the iMaterialist dataset. There are 46 categories in the original dataset, making it ideally suited to both classification and segmentation. We have segmented the input image into lower body, upper body, whole body, accessory, and shoes. The IoU metrics for U-Net and Detectron2 for segmentation results were 0.62±0.36 and 0.79±0.24 respectively. Some segmentation results for our dataset are shown in Figure 2 and Figure 3.

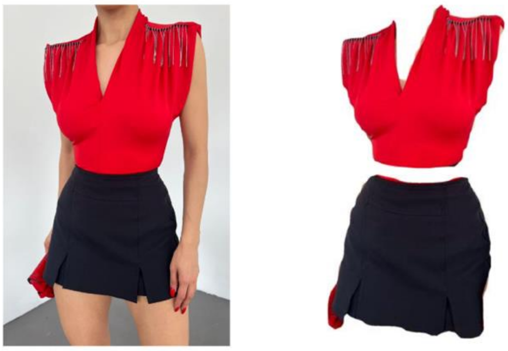


Figure 2 Segmented view of a red blouse and black skirt.

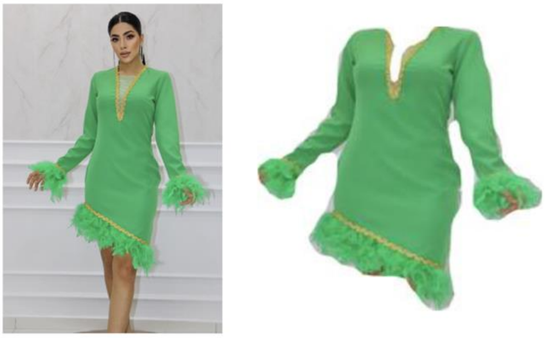


Figure 3 Advanced segmentation of a green dress with lace detailing.

2.3.Feature Extraction

In our study, we employ two distinct convolutional neural networks, ResNet50 (Residual Network [15] with 50 layers) and VGG19 [16], for the feature extraction process to compare their effectiveness in fashion image clustering. Both networks are known for their deep learning capabilities in image recognition tasks but differ significantly in architecture and performance characteristics, which can impact the clustering outcome.

ResNet50 utilizes a series of residual blocks that incorporate skip connections, allowing the model to skip one or more layers. These skip connections help to avoid the problem of vanishing gradients by allowing gradients to flow through a shortcut path during backpropagation. We use a pre-trained ResNet50 model, utilizing weights from the ImageNet dataset, which provides a robust foundational understanding of visual features across diverse categories.

For clustering, we extract features from the last fully connected layer of ResNet50, which contains 2048 deep features encapsulating high-level semantic information from the images. This layer is chosen because it provides a dense representation of the image, capturing both the abstract and detailed aspects, which is ideal for clustering based on visual similarity.

VGG19 has 19 layers, and this architecture is known for its simplicity. Like ResNet50, VGG19 is also pre-trained on the ImageNet dataset. VGG19 employs a stack of convolutional layers with small receptive fields followed by max-pooling layers, which increase the depth of the network while reducing spatial dimensions of the feature maps.

For feature extraction, we use the output from the last fully connected layer of VGG19, which, like ResNet50, provides a feature vector with a comprehensive depiction of the image. However, VGG19's feature vectors tend to be more flattened and less hierarchical compared to those from ResNet50, potentially affecting the granularity and nature of the clusters formed.

The extracted features from both ResNet50 and VGG19 are subsequently used to feature extraction before perform clustering. The choice of different architectures allows us to examine how the depth and structure of a network influence the clustering quality and accuracy. By analyzing the similar product recommendation results obtained from both models, we aim to discern which architectural traits contribute most effectively to grouping fashion images, thus providing insights into the optimal deep learning approach for such applications.

2.4.Clustering Of The Database

After the feature extraction process using ResNet50 and VGG19, we proceed with the clustering of the extracted features using the k-means clustering algorithm. K-means is chosen for its simplicity and effectiveness in grouping data into k distinct clusters based on feature similarity. The algorithm partitions the images by assigning each image to the nearest cluster center, minimizing the within-cluster variance, also known as the inertia.

Using the features extracted from both the ResNet50 and VGG19 models, k-means clustering is applied separately to each set of features. The number of clusters was chosen as 5 through the Elbow Method to finalize the clustering configuration.

The final clusters are analyzed to assess the visual and stylistic similarities within each cluster, validating the effectiveness of the chosen models and clustering techniques. By comparing the outcomes of clustering with features from ResNet50 and VGG19, we gain insights into which feature extraction method is more conducive for fashion image categorization.

2.5. Feature Reduction

Applying Principal Component Analysis (PCA) after clustering reduces the dimensionality of feature vectors extracted from images, thereby enhancing computational efficiency in subsequent operations. PCA is particularly useful when comparing a test image with database images using cosine similarity. Applying PCA to reduce the dimension from D to d results in lower computational costs for image comparisons due to the reduced complexity from $O(D)$ to $O(d)$. This enhancement makes the process more efficient and scalable, particularly useful in systems where real-time image comparisons are necessary.

3. Results and Discussion

To identify similar products, we begin by segmenting images to isolate fashion items. As we stated in previous section, we have trained Detectron2 algorithm with pre-annotated dataset iMaterialist and test on it. Since its segmentation scores are satisfactory level, we have applied our trained segmentation model to collected shop image dataset.

Next, we extracted feature vectors from each segmented product image using the architectures of ResNet50 and VGG19. By omitting the final layer of these models, we obtain a 1x2048 fully connected embedding vector for each product. To visualize these vectors, we employ t-SNE, which helps us plot the feature vectors in a two-dimensional space. A selection of 500 randomly chosen images is displayed in Figure 4, demonstrating that similar products and colors cluster together, particularly when using ResNet50 features. This pattern is also observed with VGG19, reinforcing the use of cosine similarity as a metric for identifying visually similar items.

To find products similar to a given input image, we segment fashion objects within the image, extract their features, and then search our database for similar items by calculating the cosine similarity between the feature vectors of the input product and those in the database. Figure 4 demonstrates that objects sharing similar colors and categories are clustered in proximity within the space. This observation validates the use of cosine similarity as a metric for identifying visually analogous objects in the database.

We calculated the cosine similarity for all images in our database and subsequently grouped the images into clusters. Within these clusters, we specifically identified the group that included our test images. We then calculated the cosine similarity between these test images and others in the same cluster. We also implemented Principal Component Analysis (PCA) with a range of component numbers ($n = 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024$). Afterward, we measured the pairwise cosine similarities and clustered the features post-dimensionality reduction to identify similar products within this condensed feature space. For each scenario, we recorded the computation time and evaluated performance. All experiments were conducted on a workstation equipped with an Intel Xeon W 1270 processor, 64 GB of RAM, a 1 TB hard drive, and an NVIDIA Quadro P2200 graphics card with 5 GB of dedicated memory.

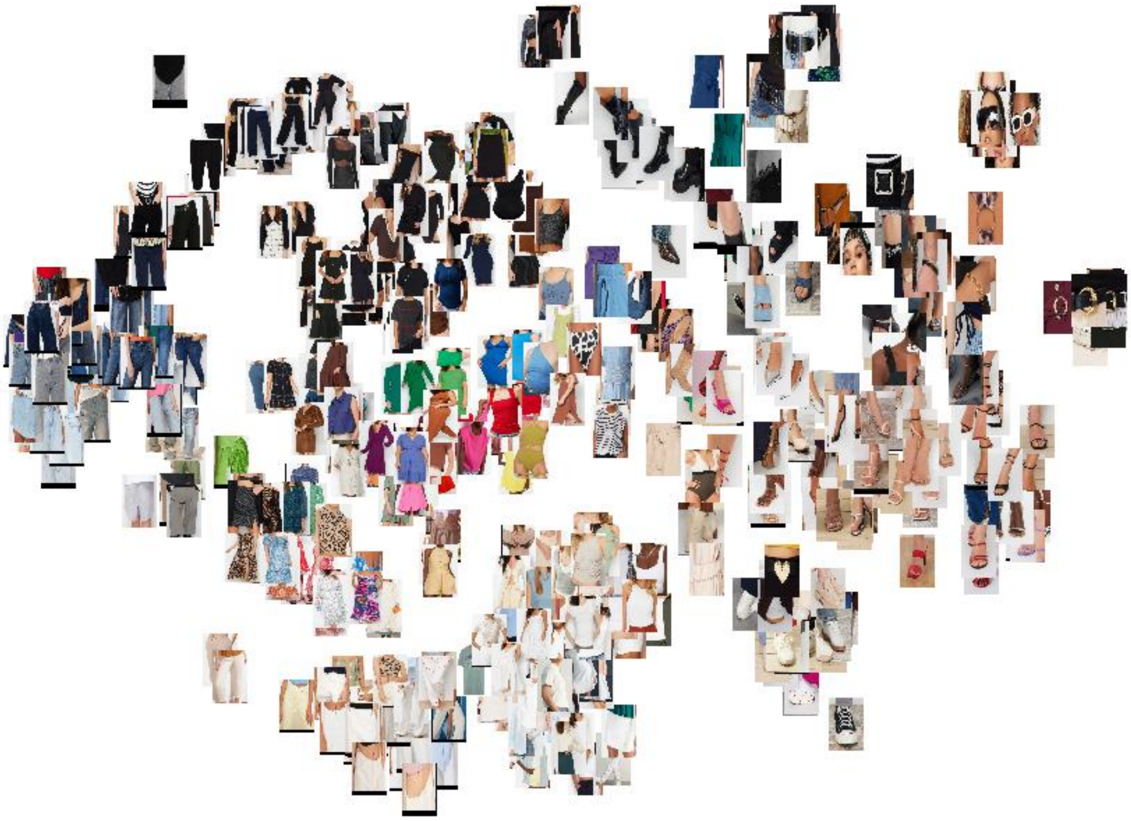


Figure 4 Visualization of 500 randomly selected product images from our dataset after feature extraction using ResNet50 and dimensionality reduction via t-SNE for 2D embedding.

Figure 5 shows the computational time versus the number of principal components. The results indicate that computation time significantly decreases as the number of principal components is reduced. Additionally, when working with a higher number of principal components, clustering further decreases the average computation time because it involves comparing fewer samples. Given the importance of speed on e-commerce web pages, even a small reduction in computation time is highly valuable.

We have multiple images taken from different angles for the same product. For each product, we determine five recommendations. Our success metric is based on how many of these five recommendations are of the same product. Considering that some products do not have five different images, resulting in a lower success rate, our results appear to be sufficiently satisfactory.

As summarized in Figure 6, the prediction performance does not change dramatically when using clustering, which is beneficial for our purposes. The graphs indicate that Resnet50 consistently outperforms VGG19 in terms of prediction accuracy. This is evident as the average number of true product recommendations is higher for Resnet50 across all principal component numbers.

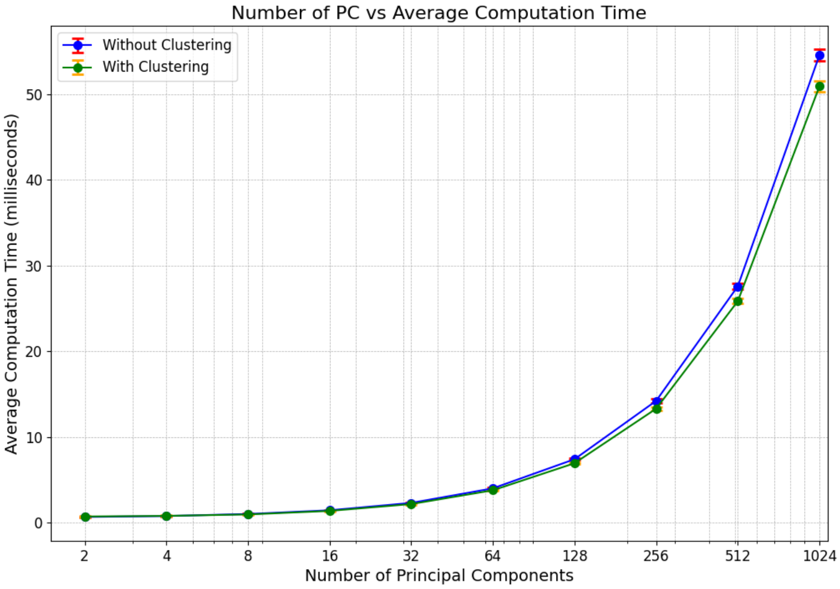
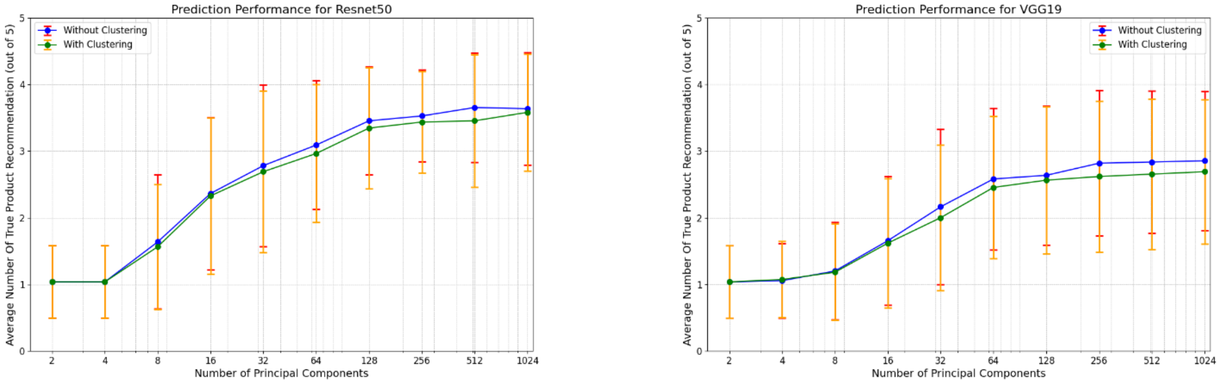


Figure 5 Average computational time for finding the top-5 similar products for a given product.

Additionally, it is important to note that clustering helps in reducing computational time by comparing fewer samples, without significantly impacting the prediction performance. This is crucial for e-commerce applications where speed is a key factor. The slight variations in the number of principal components, as shown in the graphs, demonstrate that while both models improve with more components, Resnet50 maintains a better performance level, making it a preferable choice for our image retrieval system.



(a) (b)

Figure 6 Average number of true product recommendations for a given product with feature extraction using (a) Resnet50 (b) VGG19.

We have demonstrated some visual results for a given product image in Figure 7. Considering different feature extraction methods, different numbers of PCA components, and whether clustering is applied, there are a total of 40 combinations and 200 recommendations for an input image. Therefore, we did not include all the results in the figure. To illustrate the general structure, we used both ResNet50 and VGG19 for feature extraction with n=1024 and included the top 5 recommended product images. When we look at these, we can say that the ResNet architecture yields better results in finding similar images. Therefore, we only shared the ResNet results to compare the number of PCA components. This example clearly shows that as the number of n components decreases, the performance in finding similar products decreases. All results are provided without clustering since there is only a slight decrease in performance when clustering is applied.

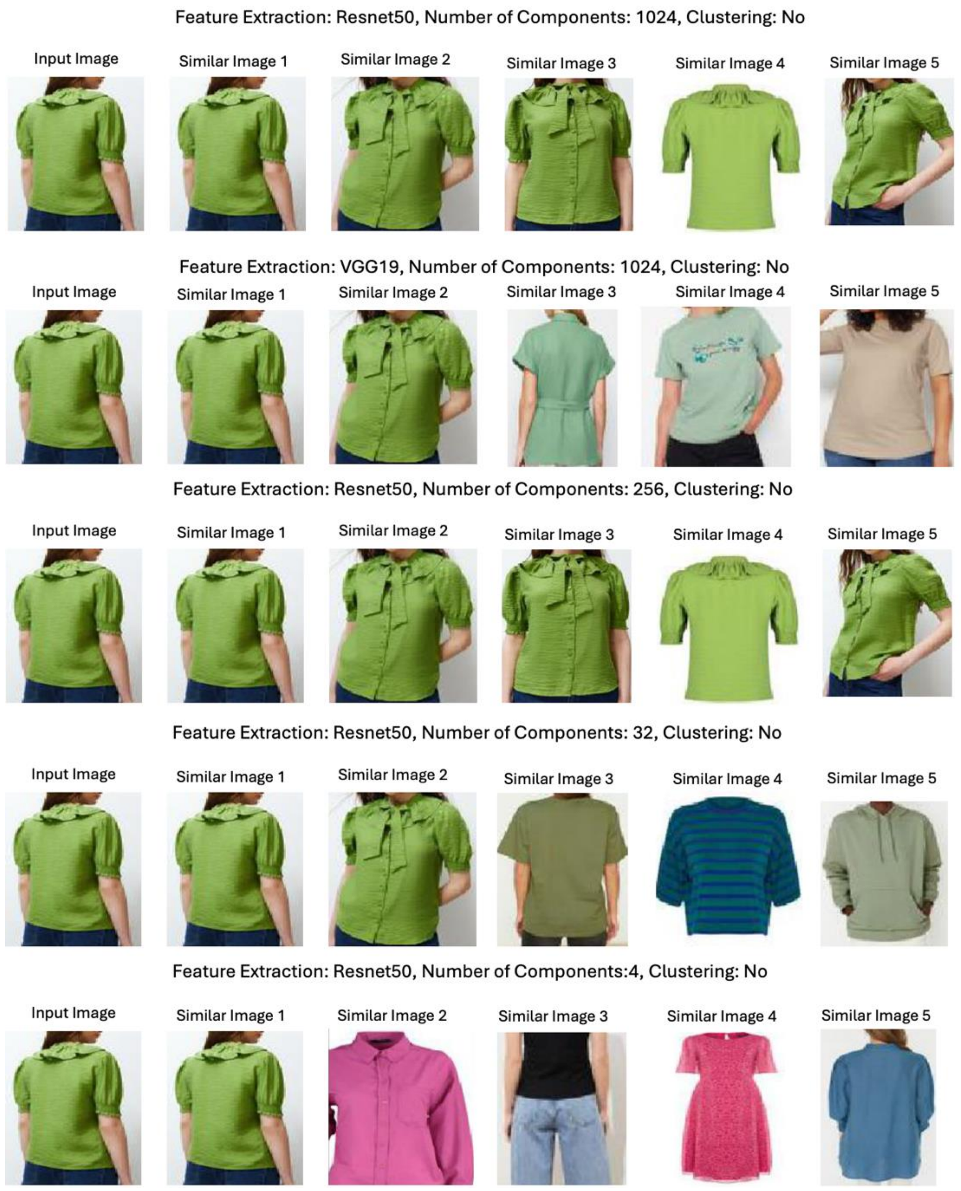


Figure 7 Selected results for various feature extraction architectures and different numbers of principal components.

4. Conclusion

In this study, we presented a novel approach to enhancing fashion image retrieval using clustering methods and Principal Component Analysis (PCA). Our findings emphasize the significant impact that feature extraction techniques have on the effectiveness of fashion image searches. By grouping similar items, we were able to make the search process faster while maintaining an acceptable level of accuracy. Notably, our experiments demonstrated that ResNet50 outperforms VGG19 in terms of prediction accuracy, particularly when combined with PCA. This combination not only speeds up the retrieval process by reducing the number of necessary comparisons but also proves to be highly beneficial for e-commerce platforms, where user satisfaction is closely tied to the speed and relevance of search results.

However, we acknowledge certain threats to the validity of our findings. The reliance on human evaluation, for instance, introduces subjectivity, as different users may perceive visual similarity differently. Furthermore, traditional evaluation metrics may not fully capture the subtle visual details that are crucial in fashion image retrieval, potentially leading to a gap between algorithmic performance and user expectations. Additionally, the dataset used in our study might not encompass the full diversity of fashion items available in the real world, which could limit the generalizability of our results.

Despite these challenges, our approach demonstrates a clear improvement in both the efficiency and accuracy of fashion image retrieval. By carefully considering these threats to validity and continuing to refine our methods, this research lays a strong foundation for future work in the field. Our approach not only optimizes the search process but also enhances the overall user experience, making it a valuable contribution to the ongoing development of intelligent fashion recommendation systems.

Ethics in Publishing

There are no ethical issues regarding the publication of this study.

Acknowledgements

This project is supported by the KOSGEB’s R&D, Product Development, and Innovation Support Program grant number 68KIB.

References

[1] S. Shirkhani, H. Mokayed, R. Saini, and H. Y. Chai, “Study of AI-Driven Fashion Recommender Systems,” *SN Comput Sci*, vol. 4, no. 5, p. 514, 2023, doi: 10.1007/s42979-023-01932-9.

[2] X. Li, X. Wang, X. He, L. Chen, J. Xiao, and T.-S. Chua, “Hierarchical fashion graph network for personalized outfit recommendation,” in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 159–168.

[3] A. Dagan, I. Guy, and S. Novgorodov, “Shop by image: characterizing visual search in e-commerce,” *Information Retrieval Journal*, vol. 26, no. 1, p. 2, 2023, doi: 10.1007/s10791-023-09418-1.

[4] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3343–3351.

[5] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[6] M. Jia *et al.*, "Fashionpedia: Ontology, segmentation, and an attribute localization dataset," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 2020, pp. 316–332.

[7] B. Kolisnik, I. Hogan, and F. Zulkernine, "Condition-CNN: A hierarchical multi-label fashion image classification model," *Expert Syst Appl*, vol. 182, p. 115195, 2021.

[8] Z.-Q. Cheng, X. Wu, Y. Liu, and X.-S. Hua, "Video2shop: Exact matching clothes in videos to online shopping images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4048–4056.

[9] N. Garcia and G. Vogiatzis, "Dress like a star: Retrieving fashion products from videos," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2293–2299.

[10] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1062–1070.

[11] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3343–3351.

[12] L. Richardson, "Beautiful soup documentation," 2007, *April*.

[13] Y. Wu, A. Kirillov, F. Massa, W. Y. Lo, and R. Girshick, "Detectron2 [www document]," *URL https://github. com/facebookresearch/detectron2 (accessed 12.12. 23)*, 2019.

[14] S. Guo *et al.*, "The imaterialist fashion attribute dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, p. 0.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.