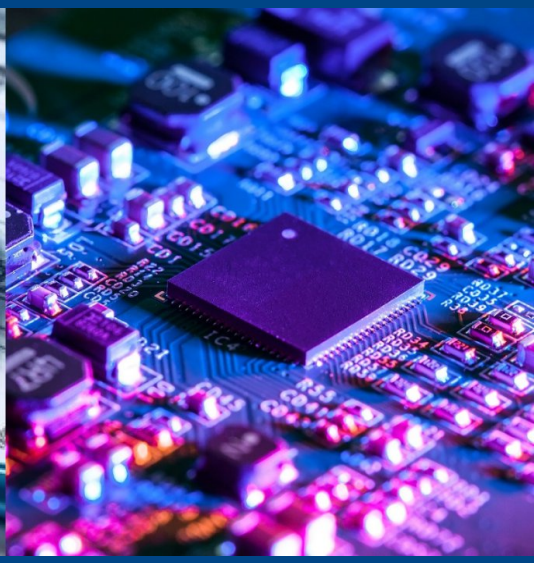
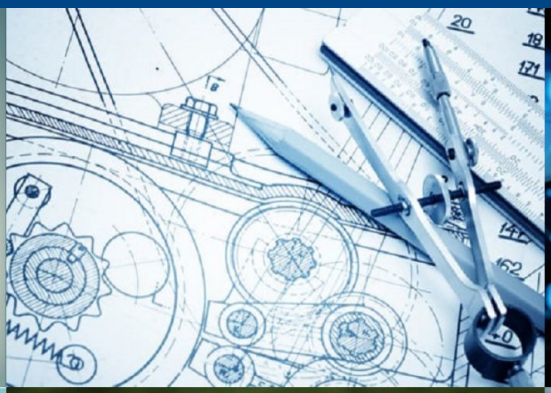
ISSN: 2582-7219

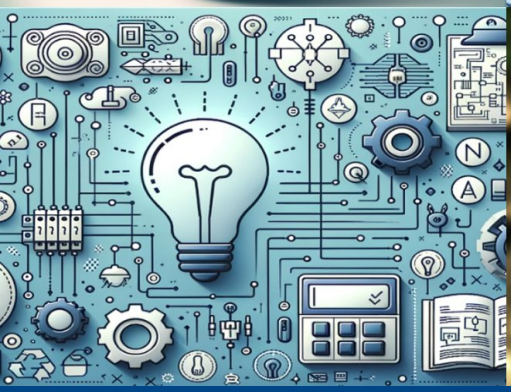




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904385

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

AI-Driven Multimodal Fashion Product Search Using CNN and Speech Recognition

Amsavalli.S, S. Mohamed Kasim

Department of Computer Applications, B.S Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India

Department of Computer Applications, B.S Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India

ABSTRACT: This paper presents a multimodal fashion product search system that integrates image-based retrieval, text-based semantic search, and multilingual speech recognition into a unified AI framework. The system leverages a pretrained ResNet50 convolutional neural network to extract 2048-dimensional visual embeddings for content-based image retrieval. Text queries are processed using TF-IDF vectorization and cosine similarity for semantic matching. Voice input is converted to text using browser-based speech recognition and processed through the same textual pipeline. To enhance personalization, a contextual AI ranking module incorporates gender, occasion, and age-based weighting into similarity scoring. Additionally, KMeans clustering is applied to deep visual embeddings to group stylistically similar fashion items. The system operates efficiently in a CPU-based environment and is evaluated on a dataset of 22,470 fashion products across 15 article categories. Experimental results demonstrate improved retrieval flexibility and contextual relevance compared to unimodal approaches.

KEYWORDS: Multimodal Search, Fashion Retrieval, ResNet50, TF-IDF, Cosine Similarity, Speech Recognition, KMeans Clustering, Contextual AI.

I. INTRODUCTION

Modern e-commerce platforms face challenges in enabling intuitive product discovery in fashion domains. Traditional keyword-based search systems often fail when users cannot accurately describe style attributes. Similarly, image-only retrieval systems ignore contextual metadata such as gender and occasion preferences. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have significantly improved visual feature extraction. Meanwhile, natural language processing techniques enable semantic understanding of textual input. However, most existing systems operate in isolation and lack unified multimodal integration. This paper proposes a unified multimodal fashion search system.

II. LITERATURE REVIEW

Fashion product retrieval has gained significant attention with the growth of e-commerce platforms. Early systems relied primarily on keyword-based search mechanisms that matched user queries with product metadata. While effective for structured queries, these systems often failed when users lacked precise fashion terminology. To address this limitation, content-based image retrieval (CBIR) techniques were introduced, leveraging convolutional neural networks (CNNs) to extract visual features from product images. Architectures such as ResNet, VGG, and Inception demonstrated strong performance in extracting discriminative features, enabling similarity-based retrieval using distance metrics such as cosine similarity.

Recent research has explored deep learning approaches for visual fashion understanding. Pretrained CNN models, particularly ResNet variants, have shown strong generalization capabilities through transfer learning. These models generate high-dimensional embeddings that represent semantic visual information. However, most image-based retrieval systems operate independently of textual metadata and lack contextual personalization. While visual similarity captures style resemblance, it does not account for user preferences such as occasion, age group, or gender-based filtering, limiting practical recommendation quality.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904385

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Parallel to image-based retrieval, text-based fashion search systems have employed vector space models and natural language processing techniques. Traditional approaches use TF-IDF representations combined with cosine similarity for efficient document matching. Although lightweight and computationally efficient, TF-IDF lacks deep semantic understanding compared to transformer- based models such as BERT. Nevertheless, TF-IDF remains suitable for CPU-based deployment scenarios due to its low computational cost and fast indexing capabilities. However, existing text-based systems generally do not integrate multimodal signals or visual embeddings in a unified architecture.

More recently, multimodal retrieval systems have emerged, combining visual and textual features for improved search flexibility. Some approaches utilize joint embedding spaces or models such as CLIP to align image and text representations. While these models demonstrate strong performance, they often require high computational resources and GPU-based inference, limiting deployment in lightweight environments. Additionally, many systems lack contextual ranking mechanisms that incorporate metadata such as age, occasion, and personalization biases. Therefore, there remains a need for a computationally efficient multimodal fashion retrieval system that integrates visual similarity, text search, speech input, and contextual personalization within a unified framework.

III. PROBLEM DEFINITION

Online fashion platforms struggle to provide intuitive and personalized search experiences. Users often rely on vague textual descriptions or uploaded images without structured query knowledge. Existing systems typically operate in isolation—either visual search or text search—leading to incomplete results. Furthermore, traditional e-commerce platforms lack contextual personalization mechanisms that consider age group, gender preferences, and occasion-specific constraints. Static ranking strategies fail to prioritize relevant items according to user intent. Additionally, many advanced AI systems require GPU infrastructure, limiting their scalability in lightweight deployments. Hence, a unified, CPU-efficient multimodal system is required to integrate visual similarity, semantic text retrieval, speech input, clustering, and contextual ranking into a single intelligent framework.

IV. SYSTEM ARCHITECTURE

The system follows a modular horizontal architecture composed of multiple integrated layers:

- Presentation Layer (Frontend Interface): Provides user interaction through image upload, text input, and voice search. Displays ranked products in an e-commerce grid layout:
- Application Layer (Flask API Server): Acts as the central controller, handling incoming requests and routing them to appropriate AI modules.
- AI processing Layer: Includes ResNet50 for visual feature extraction, TF-IDF for text processing, cosine similarity computation, KMeans clustering, and contextual AI ranking.
- Data Layer: Stores product images, metadata (gender, article type, color), and precomputed embeddings for efficient retrieval.

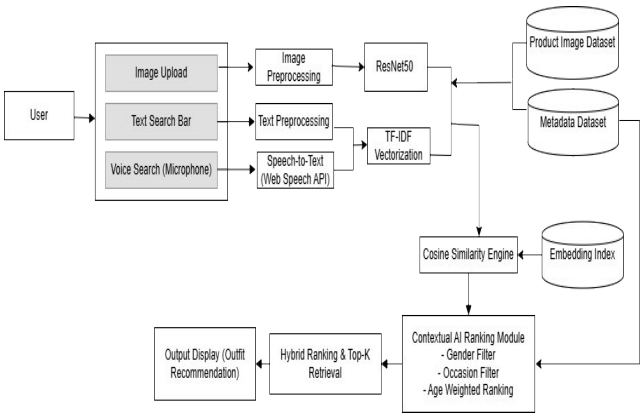


Figure 1: Architecture Diagram

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904385

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. PROPOSED METHODOLOGY

The system follows a structured pipeline from input acquisition to final ranking.

A. Data Preparation

The dataset contains 22,470 fashion products across 15 article types. Metadata fields such as gender, base color, usage, and article type are cleaned and indexed. Image embeddings are precomputed and stored for efficient similarity computation.

B. Image Feature Extraction

A pretrained ResNet50 model is used to extract 2048- dimensional embeddings from product images. The final classification layer is removed to retain deep semantic feature representations. Cosine similarity is used to compare query embeddings with stored embeddings.

C. Text Search using TF-IDF

Product metadata is converted into a combined textual representation. TF-IDF vectorization transforms textual information into numerical vectors. User queries are converted into TF-IDF vectors and matched against stored vectors using cosine similarity.

D. Voice Search Integration

Voice queries are captured using Web Speech API and converted into text. The resulting text is processed using the same TF-IDF pipeline as standard text queries.

E. Contextual AI Ranking

The ranking module applies gender-based filtering, occasion-based filtering, and age-group weighting. Instead of strict filtering, soft biasing is applied to improve personalization while preserving diversity.

F. Clustering

KMeans clustering is applied to deep visual embeddings to group stylistically similar fashion items. This supports style-based recommendation expansion.

VI. EXPERIMENTAL SETUP AND RESULTS

The dataset was divided into indexed storage and runtime query evaluation. The system was tested under CPU-only conditions to evaluate efficiency.

Evaluation metrics include:

- Precision@K
- Top-K Retrieval Accuracy
- Retrieval Latency

The multimodal system demonstrated improved retrieval flexibility compared to single-modality search. Contextual ranking improved relevance by prioritizing age-appropriate and occasion-specific items. CPU-based NumPy similarity computation ensured fast response time without GPU dependency.

VII. DISCUSSION

The proposed system successfully integrates multiple AI components into a unified framework. Unlike conventional e-commerce search engines, it supports multimodal interaction and contextual personalization. The hybrid ranking mechanism ensures both similarity-based accuracy and user-centric filtering.

A significant advantage of this architecture is its lightweight design. While transformer-based models offer deeper semantic alignment, the chosen approach balances computational efficiency with practical deployment feasibility. The integration of clustering further enhances recommendation diversity.

VIII. LIMITATIONS

Despite its strengths, the system has certain limitations. TF- IDF lacks deep semantic understanding compared to transformer-based language models. The clustering module is static and does not dynamically adapt to evolving fashion trends. Additionally, the system does not currently incorporate real-time user feedback learning. Performance is dependent on dataset quality and metadata accuracy.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IX. CONCLUSION AND FUTURE SCOPE

This research presents a multimodal AI-driven fashion product search system integrating image retrieval, text similarity, speech recognition, contextual ranking, and clustering within a CPU-efficient framework. By combining deep visual embeddings with metadata-based personalization, the system enhances search flexibility and recommendation relevance.

Future enhancements include integration of transformer- based semantic models, deployment of joint image-text embedding architectures such as CLIP, incorporation of user feedback learning mechanisms, and large-scale cloud deployment.

REFERENCES

[1] **L. Sivaranjani, B. R. A, S. K. Rachamadugu, M. Sakthivel, B. V. S. Reddy, and S. Depuru**, "Fashion Recommendation System Using Machine Learning," 2023.
[2] **A. Sheeba, S. S. H., and S. S.,** "Voice Enabled E-Commerce Website for Visually Impaired People and Non-Disabled Users," Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 24), 2022.
[3] **D. Muppalla, P. Pranjali, R. Mishra, S. B. Agarwal, and A. M.,** "AI-Powered Visual Search and Virtual Try-On for E-Commerce," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 13, no. IV, Apr. 2025.
[4] **S. T. Jami, G. R. Karri, H. U. Rahman, S. S. Merupo, and V. S. D. Batchu,** "Style Matcher: A Deep Learning Framework for Visual Fashion Matching Using ResNet50," World Journal of Advanced Engineering Technology and Sciences, vol. 15, no. 01, pp. 1710–1721, 2025.
[5] **Z. F. S. Shariff,** "Product Matching for E-commerce Platform Based on Text and Image Similarity Using Deep Neural Network Architecture," MSc Research Project, National College of Ireland, 2022.
[6] **A. Khan, A. Imdad, K. U. Safi, and F. Ahmed**, "Exploring the Impact of Voice Search and Voice Commerce on Consumer Shopping Habits and Brand Interactions," 2025.
[7] **Y. S. Yaswanthraj, M. S. Mohithra, P. K. Nadar, and S. S.,** "AI- Based Outfit Recommendation System," 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com