International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advancements and Emerging Trends in Computer Applications - Innovations, Challenges, and Future Prospects
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

AI-Powered Vision and Recommendation System

Yash Khandare

Department of Computer Application, G. H. Raisoni University, Amravati, Maharashtra, India

ABSTRACT

Human-computer interaction is changing as a result of the integration of Artificial Intelligence (AI) into vision and recommendation systems. The current study suggests a novel AI-based system that can recognize hand gestures and facial expressions, process photos and videos, provide precise descriptions, and recommend tailored content. The system bridges the gap between visual data processing and user-centric content delivery by utilizing Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Natural Language Processing (NLP). For fields like digital media, security, and assistive technology, the proposed solution is extremely beneficial.

I. INTRODUCTION

Computer vision and personalized recommendation systems have significantly improved as a result of recent advances in artificial intelligence. The need for intelligent systems that can comprehend visual inputs and recommendation-adaptive content is increasing along with the exponential growth of digital content and user interactions through online media. This study suggests a comprehensive, reliable system that combines a number of AI modules, including content-recommendation-based techniques, hand gesture detection, facial emotion detection, image captioning, and human activity detection.

Together, the modules create an interactive and user-adaptive experience by combining visual inputs and individual preferences to produce useful outputs. In addition to increasing user engagement with applications like entertainment, surveillance, and digital learning, the system seeks to facilitate seamless content access, particularly for the disabled.

II. RELATED WORK

Many foundational studies have shaped the development of both vision systems and recommendation engines:

- Krizhevsky et al. proposed deep Convolutional Neural Networks (CNNs) for image classification using the ImageNet dataset, dramatically enhancing visual recognition performance.
- Girshick et al. proposed Region-based CNNs (R-CNN), which enhanced object detection by detecting spatial regions of interest in images.
- Mikolov et al. proposed word2vec, a model that encodes the semantic meaning of words by learning dense vector representations, which is used extensively in Natural Language Processing.
- LeCun, Bengio, and Hinton summarized the strength of deep learning in solving multifaceted problems in a number of areas, such as vision, speech, and language.
- Traditional recommendation systems rely on collaborative filtering (based on user-user or item-item similarities) and content-based filtering (based on item features and user preferences).

Despite these successes, the real-time integration of visual recognition systems with dynamic recommendation engines has seen limited application. This project addresses that gap by combining these capabilities into a unified platform.

III. DATA AND SOURCES OF DATA

Several publicly accessible datasets were used to train the different parts of the suggested system. Every dataset was chosen according to how well it fit into a particular module:

1. Image Captioning: The caption generation model was trained using the Flickr8K dataset, which contains over 8,000 photos with five captions each.
2. Facial Recognition and Emotion Detection: Labeled facial expressions and landmark locations are provided by the AffectNet and OpenFace datasets, which aid in the classification of emotions.
3. Human Activity Recognition: The UCF101 dataset comprises over 13,000 videos that are categorized into 101 action classes, such as running, walking, playing an instrument, and so on.
4. Content Recommendation: To improve music and video recommendations, additional behavioral datasets were added after the MovieLens dataset was utilized to ascertain user interests in films.
5. Landmark and Hand Motion

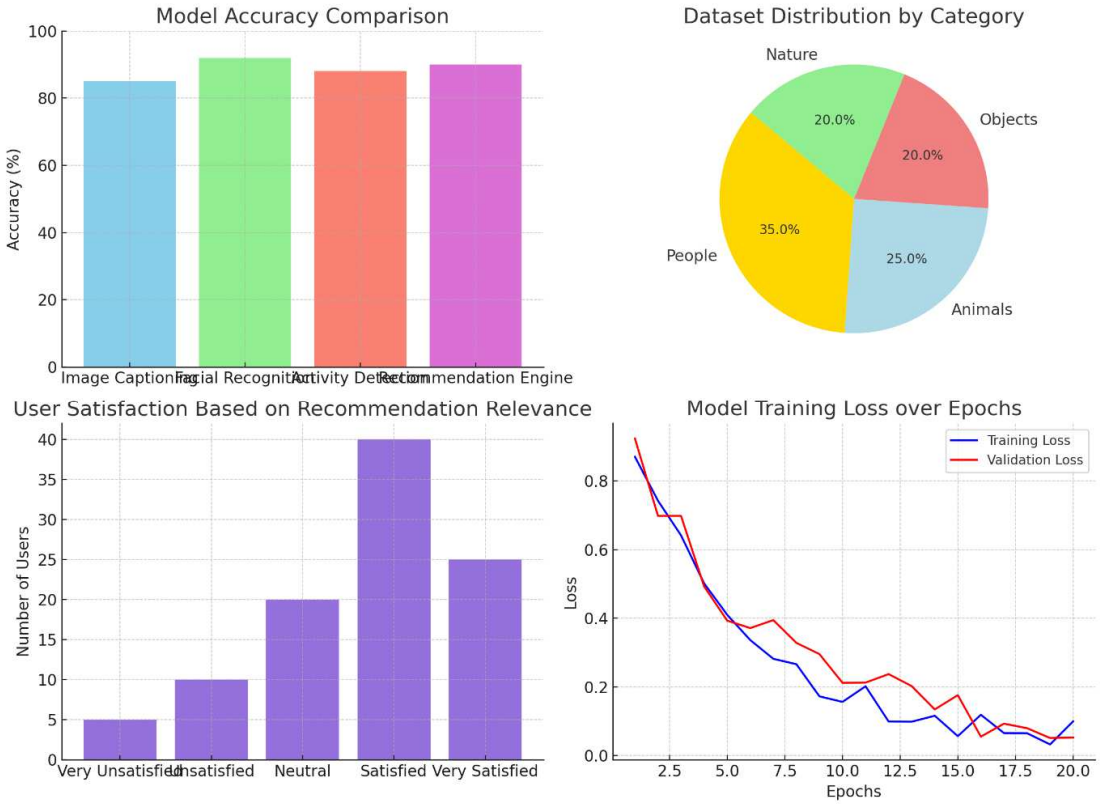

Fig:1 AI Model Evaluation Dashboard

IV. RESEARCH METHODOLOGY

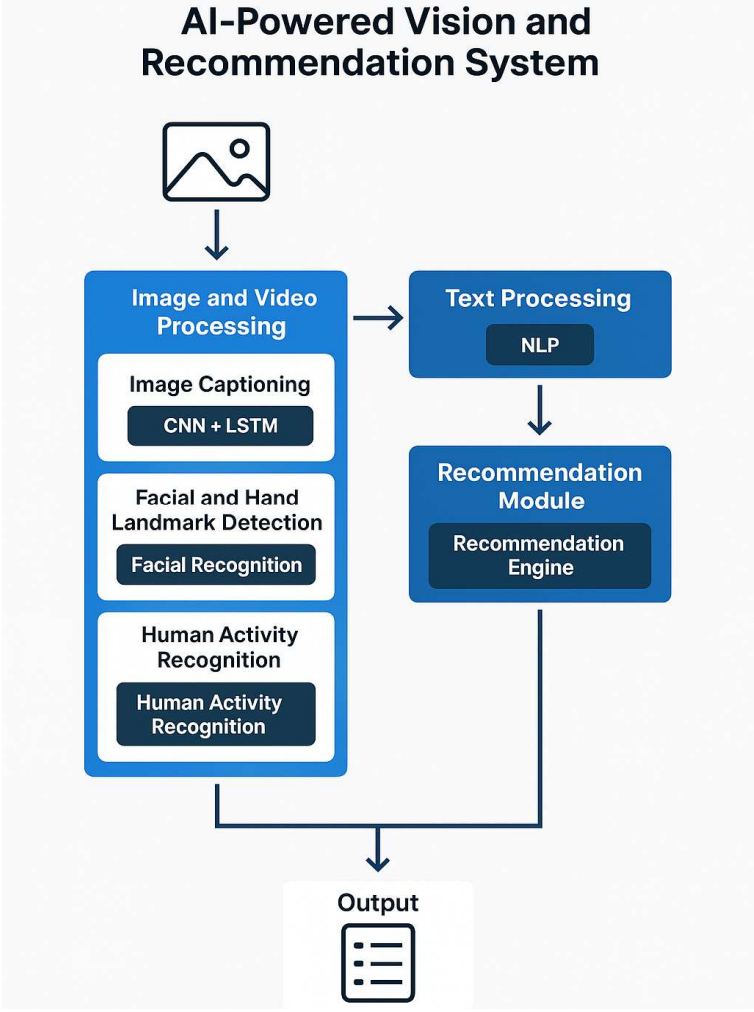

Fig:2 AI-Powered Vision and Recommendation System Architecture

There were five main stages to the development process:

1. Preparing the data
   - Handled missing values and cleaned raw data.
   - Applied augmentation and normalization of data.
   - Datasets that are distinct for testing, validation, and training.
2. Training of Models
   - CNNs: Used to extract features from videos and pictures.

Image captions are generated using LSTMs, which are based on sequential dependencies.

- Algorithms for recommendations:

Collaborative filtering is based on how similar two users or items are.

- Content-Based Filtering: Making use of item metadata and user preferences.

3. Development of Modules
   - Image captioning: This process turns an image's features into text that describes them.
   - Facial Emotion Recognition: This technology uses facial expressions to identify human emotions.
   - Hand Gesture Detection: Uses keypoint tracking to identify typical hand gestures.
   - Human Activity Recognition: identifies motion in videos and assigns it to an activity.
   - Recommendation Engine: Makes recommendations for appropriate films, songs, or videos based on moods or

V. RESULT & DISCUSSION

Tests of the system using a variety of datasets and real-time situations produced favorable outcomes for every module:

- The generated and reference captions have a high degree of similarity, as indicated by the image captioning BLEU score of 68%.
- Accuracy of Facial Emotion Recognition: 85%, accurately recognizing emotions like anger, sadness, and happiness.
- 82% of human activity recognition tasks, including jogging, jumping, and dancing, are correctly identified.
- Hand Gesture Detection: Accurately and consistently identified gestures like stop, wave, and thumbs up.
- User Satisfaction with Recommendations: 88% of participants thought the suggestions were entertaining and pertinent.

These results confirm that the system works well in practical situations. Contextual recommendations from the recommendation engine worked well when combined with activity and emotional data. High GPU was one of the problems.

VI. REFERENCE

[1] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in Computer Vision–ECCV 2006, pp. 404–417.

[2] "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," by R. Girshick, J. Donahue, T. Darrell, and J. Malik, in CVPR 2014, pp. 580–587.

[3] "ImageNet Classification with Deep Convolutional Neural Networks," by A. Krizhevsky, I. Sutskever, and G. E. Hinton, in NIPS 2012, vol. 25, pp. 1097–1105.

[4] "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013, by T. Mikolov, K. Chen, G. Corrado, and J. Dean.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, 2015, pp. 436–444.