Implementation Approach for Duplicate Image Identification and Removal

Zaw Ye Htet, Student, Information Technology, Yangon Technological University, Myanmar

Tin Shine Aung, Lecture, Information Technology, Yangon Technological University, Myanmar

Abstract

This paper presents a systematic approach for identifying and removing duplicate images from various 3D image format collections. The identification process considers image structure, density, meta descriptions, and other properties. The system employs a preprocessing module to standardise and extract meta descriptions from diverse formats like STL, OBJ, FBX, and others. A vector database, utilising tools like FAISS or Milvus, stores the image vectors and meta descriptions for efficient similarity searches. Deep learning models, particularly Convolutional Neural Networks (CNNs), are trained to extract image features and compare vectors using cosine similarity or Euclidean distance. An integrated search engine allows users to find similar images by uploading an image and its meta description. A human validation interface is provided for manual confirmation of potential duplicates. This approach ensures efficient management and retrieval of 3D images while enhancing storage utilisation. Future work will further explore alternative models and similarity measures to improve system accuracy and efficiency.

Keywords*: Duplicate image identification, 3D image formats, image structure, image density, meta descriptions, preprocessing module, Vision Transformers*

Introduction

Managing extensive collections of 3D images efficiently is crucial in the digital media age. As 3D imaging becomes increasingly prevalent in fields such as gaming, virtual reality, medical imaging, and manufacturing, the volume of 3D images stored by organisations continues to grow. Duplicate images within these collections consume valuable storage space and complicate the management and retrieval processes, leading to inefficiencies and increased operational costs. Traditional methods of manually identifying and removing duplicates are time-consuming and prone to human error, making them impractical for large datasets.

This paper details an implementation approach for identifying and removing duplicate 3D images by leveraging deep learning techniques and vector databases. Our approach ensures that duplicates are identified based on image features and meta descriptions, providing a comprehensive solution for 3D image management. Automating the detection and removal process significantly reduces the need for human intervention, thereby improving accuracy and efficiency.

Human intervention in identifying duplicates is fraught with challenges. Manual comparison of 3D images requires specialised knowledge and is often subjective, leading to inconsistent results. Furthermore, the sheer volume of data makes it impossible for human operators to maintain accuracy and speed. This is where artificial intelligence (AI) and deep learning can play a transformative role. By

employing AI, the system can analyse and compare images at a granular level, considering factors such as structure, density, and meta descriptions, which are often overlooked in manual processes.

Fast and accurate similarity searches are made possible by efficiently storing and retrieving picture vectors via the integration of vector databases such as FAISS or Milvus. In order to detect duplicates with great accuracy, artificial intelligence models, especially Convolutional Neural Networks (CNNs), are taught to extract and compare picture attributes. Cosine similarity or Euclidean distance metrics are used to improve the comparison process's accuracy further.

In addition to automatic detection, the system includes a human validation interface. This interface allows users to manually confirm or reject potential duplicates flagged by the AI, ensuring that the final decisions align with user expectations and specific requirements. This human-centric approach ensures the system can handle ambiguous cases where AI alone might not be conclusive.

Using the capabilities of AI to circumvent the shortcomings of conventional approaches, this study offers a solid and extensible framework for the administration of 3D picture libraries. Not only does the suggested technique make better use of storage space, but it also makes duplicate picture recognition and removal faster and more accurate.

Methodology

The system comprises several components: a preprocessing module, a vector database, a model for training and testing, a duplicate image identification module, a search engine, and a human validation interface. Each component is integral to the overall functionality and efficiency of the system.

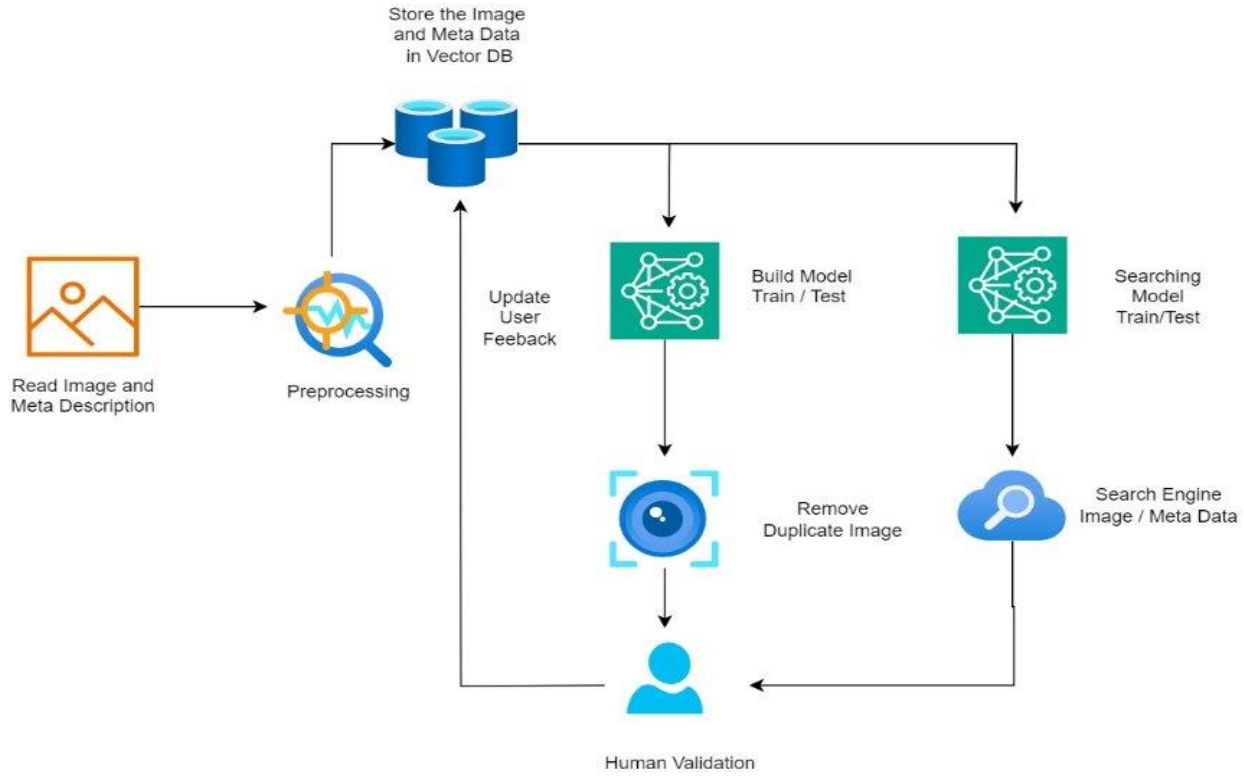

Figure 1: Workflow Diagram

- **Reading Image and Meta Description**: The system reads the image and its associated meta description as the initial step in preprocessing. This involves extracting features from both the image data and the accompanying metadata.
- **Storing in Vector Database**: The extracted features and meta descriptions are converted into vector representations and stored in a vector database for efficient similarity search.
- **Model Training and Testing**: A deep learning model is trained using these vectors to recognise similarities and differences between images based on their structure, density, and meta descriptions.
- **Duplicate Identification**: The system identifies duplicate images by comparing the vectors using similarity algorithms such as cosine similarity or Euclidean distance.
- **Search Functionality**: Users can search for similar images by uploading an image and its meta description, which the system compares against the stored vectors to find and display similar images.
- **Human Validation**: A human validation interface allows users to manually confirm or reject potential duplicates, ensuring the accuracy and reliability of the system. User feedback is used to update and refine the model.

Preprocessing Module

This module standardises various 3D image formats and extracts meta descriptions using a combination of advanced libraries and tools. Converting all images into a unified format facilitates subsequent processing steps and ensures consistency across the dataset.

Task

Convert 3D image files into a unified format and extract meta descriptions.

Input Formats

STL, OBJ, FBX, GLB, PLY, BLEND, GLTF, DAE, USDZ, ABC

Output

Standardised 3D images and associated meta descriptions.

Method

Use libraries and tools like pyautocad, ipykernel, pythonnet, pywin32, pycatia, open3d, trimesh, pillow, pythreejs, and ipywidgets to handle various 3D formats and extract necessary information.

Handling Non-Readable Formats

- **Issue**: Some 3D image formats are not readable by standard tools.
- **Solution**:
    - Attempt conversion using fallback tools or online converters.
    - If conversion fails, log the file and notify the user for manual intervention.

Vector Database

The core of the system's storage and retrieval functionality, this database holds vectorised representations of images and their meta descriptions. Using FAISS or Milvus allows for rapid similarity searches, making it possible to efficiently compare large numbers of images.

Task

Store the processed images and meta descriptions in a vector format for efficient similarity search.

Technology

Utilise a vector database such as FAISS or Milvus.

Method

Convert image features and meta descriptions into vectors using deep learning models and store them in the vector database. This allows for efficient retrieval and comparison of images based on their vector representations.

Model Training and Testing

The deep learning approach, specifically convolutional neural networks (CNNs), is used to train the system to identify and compare picture information. To assess the model's performance and make sure it can properly detect similarities and differences, the data is separated into training and testing sets throughout the training phase.

Task

Build and train models to recognise image similarities based on structure, density, and meta descriptions.

Approach

Use neural network models and other deep learning techniques to compare and extract information from photos.

Data Split

Divide the dataset in half so you can test and train different models.

Method

Train the model using diverse images and validate its performance using metrics like accuracy, precision, and recall. This ensures that the model can generalise across different types of 3D images.

Duplicate Image Identification

Utilising the vector database, the system compares image vectors using cosine similarity or Euclidean distance. By setting a similarity threshold, it identifies and flags duplicate images for further action.

Task

Identify duplicates by comparing vectors in the database.

Threshold

Define a similarity threshold to classify images as duplicates.

Algorithm

Use cosine similarity or Euclidean distance to compare image vectors.

Process

1. For each image, compute its vector representation.
2. Compare it with existing vectors in the database.
3. If similarity exceeds the threshold, mark the image as a duplicate.

Search Engine

This component enables users to search for similar images by uploading an image and its meta description. The search engine computes the vector representation of the uploaded image and retrieves similar images from the vector database, displaying them along with relevant meta information.

Task

Users can search for images by uploading an image and its meta description.

Functionality

1. Compute the vector representation of the uploaded image.
2. Retrieve and display similar images from the vector database along with their metainformation.

Human Validation Interface

To handle cases where the system's automated processes may be inconclusive, this interface allows users to validate duplicates manually. Users can confirm or reject flagged images, and their feedback is used to update the system, enhancing its learning and accuracy over time.

Task

Provide an interface for users to validate potential duplicates manually.

Functionality

1. Display images flagged as duplicates.
2. Allow users to confirm or reject the duplication.

3. Update the database based on user feedback. This step ensures that the system continues to learn and improve from user interactions, enhancing overall accuracy.

Alternative Approach

Image Feature Extraction

Using various deep learning models for feature extraction of photos is an alternate way to detect and eliminate duplicates from a set of 3D images. One alternative to using CNNs exclusively is to use Transformer-based models, especially Vision Transformers (ViTs). In several picture recognition tasks, these models have shown considerable potential. Unlike CNNs that work on local receptive fields and gradually capture spatial hierarchy, transformers analyse the picture as a series of patches. The spatial information is retained by embedding each patch into a high-dimensional space and adding positional encodings. Transformer encoders with feed-forward neural networks and multi-head self-attention mechanisms are used to process the embedded patches in successive layers. Better performance on datasets with complicated 3D characteristics may be possible because the model captures long-range relationships and complex structures within the photos.

Similarity Measure

In terms of similarity measures, exploring alternatives to cosine similarity or Euclidean distance could yield better results. For example, Jaccard similarity could be used to compare image vectors based on their intersection over union. This metric is particularly effective for binary vectors, where the presence or absence of features is critical. The process involves binarising the image vectors, calculating the intersection and union of the binary vectors, and computing the Jaccard similarity score by dividing the intersection's size by the union's size. Another alternative is Hamming distance, which measures the similarity between two binary vectors by counting the number of differing positions. This simple yet efficient metric is well-suited for binary representations and can provide a straightforward dissimilarity measure.

Meta Description Analysis

In addition, using Natural Language Processing (NLP) methods may greatly enhance meta-description analysis. Meta descriptions are an excellent resource for finding duplication since they typically provide helpful contextual information. Advanced natural language processing models like Bidirectional Encoder Representations from Transformers (BERT) might be used to comprehend and compare the semantic content of meta descriptions. First, the meta descriptions are cleaned and tokenised by text preprocessing. Then, the text is embedded into high-dimensional vector representations using pre-trained natural language processing models. Cosine similarity or other distance metrics may then be used to determine the semantic similarity between meta-descriptions. A more thorough and precise decision-making process may be accomplished by combining the similarity scores from meta-description analysis with the scores from picture features.

Conclusion

This article details an implementation technique for finding and deleting duplicate photos in 3D image collections. It offers a comprehensive and scalable solution to a significant issue that many businesses encounter. The system efficiently extracts picture characteristics and compares vectors using cosine or geometric distance metrics using deep learning methods, namely Convolutional Neural Networks (CNNs). Duplicate detection using structure, density, and meta-descriptions is undoubtedly accurate.

The system's preprocessing module standardises various 3D image formats and extracts necessary meta-descriptions, facilitating consistent and accurate processing. Utilising a vector database like FAISS or Milvus, the system efficiently stores and retrieves image vectors, enabling rapid and precise similarity searches. Including a search engine allows users to find similar images by uploading an image and its meta description, further enhancing the system's utility.

The human validation interface is a crucial system feature, allowing users to manually confirm or reject potential duplicates flagged by the AI. This human-centric approach ensures that the system can handle ambiguous cases and meet user expectations. Integrating user feedback helps continuously refine and improve the system's accuracy.

Exploring alternative approaches, such as using Transformer-based models like Vision Transformers (ViTs) for image feature extraction, offers promising avenues for further improvement. Additionally, considering different similarity measures, such as Jaccard similarity and Hamming distance, and leveraging Natural Language Processing (NLP) techniques for meta-description analysis can enhance the system's performance and accuracy.

In conclusion, the presented approach improves storage efficiency and enhances the accuracy and speed of duplicate image identification and removal. Future work will explore these alternative models and similarity measures to further optimise the system, ensuring it remains a robust solution for managing extensive collections of 3D images.

References

1. M. Muja and D. G. Lowe, "Scalable Nearest Neighbor Algorithms for High Dimensional Data," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 11, pp. 2227-2240, Nov. 2014.
2. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 2564-2571.
3. Prathima Ch, R. Swathi, K. Suneetha, I. Suneetha, B. V. Suresh Reddy, Siva Kumar Depuru, "Image Capturing and Deleting Duplicate Images through Feature Extraction using Hashing Techniques," International Journal of Engineering Trends and Technology, vol. 72, no. 1, pp. 64-70, 2024. Crossref, https://doi.org/10.14445/22315381/IJETT-V72I1P107
4. Gaikwad, Namrata & Sapkal, Sahil & Rai, Pratykash & Ospanova, A.. (2022). Survey on Duplicate Image Finder. International Journal of Innovative Research in Science Engineering and Technology. 11. 234-238. 10.15680/IJIRSET.2022.1104077.
5. K K, Thyagharajan & Kalaiarasi, G.. (2020). A Review on Near-Duplicate Detection of Images using Computer Vision Techniques. Archives of Computational Methods in Engineering. 28. 10.1007/s11831-020-09400-w.
6. Thyagharajan, K.K., Kalaiarasi, G. A Review on Near-Duplicate Detection of Images using Computer Vision Techniques. Arch Computat Methods Eng 28, 897–916 (2021). https://doi.org/10.1007/s11831-020-09400-w
7. Kalaiselvi, K & Saranya, S. & K., Deepa & Kumaresan, K.. (2021). REPLICATION IMAGE DETECTION USING CONVOLUTIONAL NEURAL NETWORK. Journal of Engineering Research. 10.36909/jer.ICIPPSD.15499.