Pinnacle Academic Press
Proceedings Series

Vol. 2 2025

Article **Open Access**

Research on Application of Big Data Mining and Analysis in Image Processing

Jialu Yan [1,*]

[1] Decoded Advertising, New York, 10005, USA
* Correspondence: Jialu Yan, Decoded Advertising, New York, 10005, USA

Abstract: Through the mining and analysis of big data, big data technology has demonstrated excellent efficiency and intelligent capabilities in image processing. This article explores the application of big data in the field of graphics, elaborates on the basic concepts and key technologies of data mining and analysis, and explains its operational processes in graphic feature screening, classification, search, and other aspects. Analyzed the application path of big data in visual presentation, quality assessment, content parsing, and semantic extraction of graphics. Research has found that big data technology is effective in addressing the challenges of diversity and large-scale data in image processing, providing new solutions in multiple industries such as smart city construction, medical image diagnosis, and security monitoring.

Keywords: big data mining; big data analysis; image processing; feature extraction; image classification

Received: 16 April 2025
Revised: 30 April 2025
Accepted: 10 May 2025
Published: 12 June 2025

1. Introduction

In today's rapidly developing information age, the speed at which image data is generated and accumulated has increased exponentially, rendering traditional image processing methods inadequate. In the face of this situation, the integration of big data technology has brought a turning point to the field of image processing. By relying on efficient data collection, storage, and analysis capabilities, big data mining technology can quickly identify the core attributes of images. Moreover, big data analysis technology enables more accurate classification and semantic understanding by deeply interpreting various dimensions of image data. How to efficiently integrate big data mining and analysis techniques to optimize image processing still faces many challenges. This study aims to analyze the application path and advantages of big data technology in the field of image processing, evaluate key technologies in image feature extraction, classification, search, quality evaluation, and provide reference for theoretical exploration of image processing.

2. Concepts Related to Big Data

2.1. Big Data Mining Technology

Big data mining technology plays an important role in processing complex data. The key is to use mathematical models, statistical analysis, and intelligent algorithms to analyze complex data, uncover hidden patterns, and assist in decision-making. The specific process typically includes several standard stages, often illustrated in technical diagrams. The technology can generally be divided into four main stages: data preprocessing, pat-

tern discovery, knowledge representation, and result evaluation. In the data preprocessing stage, the main focus is on cleaning, transforming, and dimensionality reducing the raw data to better adapt to mining needs. In the pattern discovery stage, potential patterns in the data are explored through methods such as classification, clustering, association analysis, and sequence analysis [1]. The knowledge representation and evaluation stage summarizes and confirms the effectiveness of mining results through visualization tools and model optimization methods. In the field of visual information processing, relying on its outstanding data processing capabilities, large-scale data mining technology is widely deployed in image feature screening, recognition pattern construction, and image content semantic parsing. For example, convolutional neural networks (CNNs) under deep learning architecture have been confirmed to have significant effectiveness in processing large-scale image data classification and key feature extraction. The parallel data processing technology of distributed computing systems such as Spark has improved the efficiency of processing massive amounts of image data. With the enhancement of computing performance and continuous improvement of algorithm technology, large-scale data mining techniques are expected to bring more technological innovations to the image information processing industry [2]. In the application of data mining techniques, clustering algorithms play a crucial role as they help uncover hidden patterns in datasets. Taking the commonly used *K*-means clustering algorithm as an example, its goal is to divide the dataset $X = \{x_1, x_2, \ldots, x_n\}$ into K clusters $C = \{C_1, C_2, \ldots, C_k\}$, by minimizing the following objective function:

$$J = \sum_{k=1}^{K} \sum_{x \in C_k} ||x - \mu k||^2 \quad (1)$$

In formula (1), μk is the center of cluster C_k, $\mu k = \frac{1}{|Ck|} \sum_{x \in Ck} x$. $\|x - \mu k\|^2$ the squared Euclidean distance between data point x and cluster center μk. By repeatedly optimizing the objective function using big data mining methods, it becomes possible to intelligently screen and summarize image groups with similar features from vast amounts of image data, thus laying a solid foundation for feature extraction and pattern recognition. By integrating distributed computing technologies such as Spark MLlib, this clustering approach enhances the efficiency of processing large-scale datasets and provides strong technical support for key processes such as image segmentation and category partitioning in the field of image processing (Figure 1).

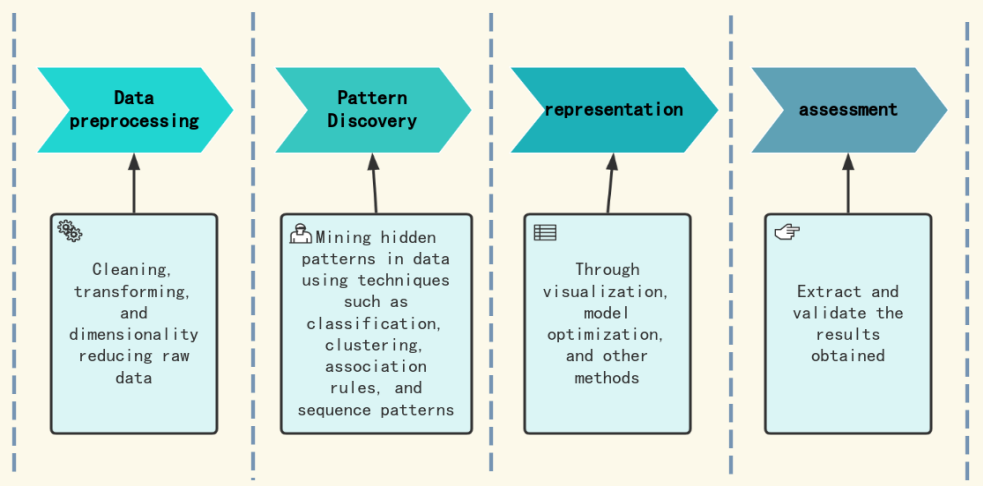

Figure 1. Big Data Mining.

2.2. Big Data Analysis Technology

Big data analysis technology refers to the techniques of using deep analytical methods to extract key information from massive amounts of data. The core elements include data modeling, statistical analysis, machine learning, and data visualization techniques,

as shown in Figure 2. Data modeling is achieved through the abstraction and mathematical representation of the inherent architecture of data. Statistical analysis uses mathematical statistical methods to explore patterns in data. Machine learning utilizes strategies such as supervised learning, unsupervised learning, and reinforcement learning to uncover hidden patterns in data [3]. Data visualization presents analysis results in an easily interpretable form, which helps people better understand and make decisions. In image processing, big data analysis technology plays an important role. By analyzing factors affecting image quality, big data analysis technology can help identify and improve key constraints in the image processing workflow. In terms of image content analysis, this technology utilizes semantic segmentation and image description generation algorithms to convert image information into organized data formats [4]. By integrating natural language processing techniques, deep semantic mining and cross modal analysis of images can be achieved. Faced with the continuous growth of image data volume and the constantly changing demand for analysis, big data analysis technology has injected strong impetus into the field of image processing [4].

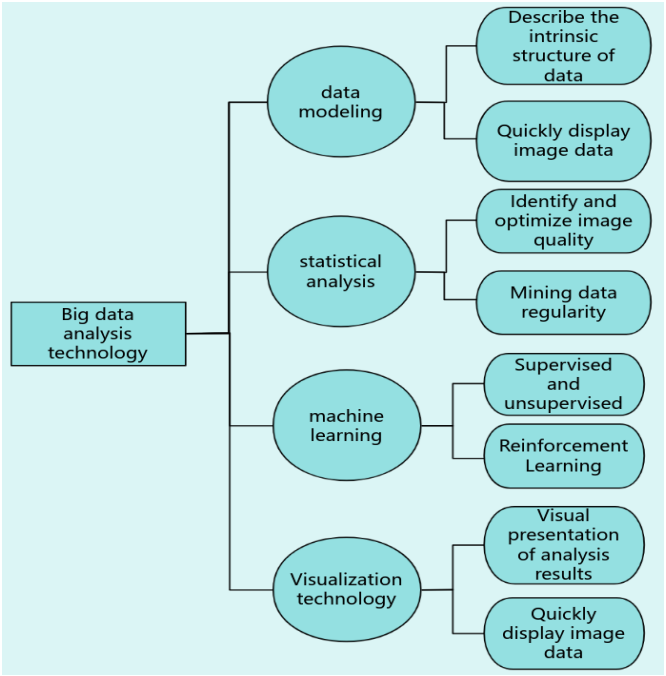

Figure 2. Big Data Analysis Techniques.

3. Application of Big Data Mining in Image Processing

3.1. Image Feature Extraction Based on Big Data Mining

In the field of image processing, extracting key features of an image is an important step that involves selecting key information from the image for in-depth analysis [5]. Relying on big data mining technology, combined with advanced deep learning algorithms and powerful distributed computing architecture, we are able to achieve efficient extraction of image features. By adopting the image feature extraction technique of Convolutional Neural Networks (CNN), features are gradually extracted through layers of stacked convolution and pooling layers. By using dimensionality reduction methods such as principal component analysis (PCA) and linear discriminant analysis (LDA), core feature information can be extracted from high-dimensional data. For a given image $X \in R^{m \times n}$, features can be extracted through convolution operations:

$$F_{i,j} = \sum_{p=-k}^{k} \sum_{q=-k}^{k} W_{p,q} \cdot X_{i+p,j+q} + b \quad (2)$$

In formula (2), $F_{i,j}$ are feature maps, $W_{p,q}$ are convolution kernel weights, b is a bias term, and k is half the size of the convolution kernel. In a big data environment, implementing parallel operations for feature extraction using distributed architectures such as Spark or Hadoop can improve processing speed. For example, cutting the image into multiple parts, distributing them to various nodes for parallel convolution operations, and finally merging the computational results to overcome the performance barriers of large-scale image feature extraction.

3.2. Image Classification Based on Big Data Mining

Image classification is a key step in image processing, aiming to categorize images based on their characteristics according to a predefined classification system. Relying on big data mining methods, combined with deep learning algorithms and distributed computing technology, it provides strong technical support for processing massive image recognition work. Among numerous classification algorithms, such as Support Vector Machine (SVM), Random Forest (RF), and Convolutional Neural Network (CNN), deep learning architectures have been widely applied and promoted due to their outstanding performance. Taking multi classification tasks as an example, for the input image dataset $X = \{x_1, x_2, \ldots, x_n\}$ and corresponding labels, $Y = \{y_1, y_2, \ldots, y_n\}$, the goal of the classification model is to find the mapping function f: X → Y. The loss function of neural network models usually adopts cross entropy loss:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\, log(\hat{y}_{i,j}) \tag{3}$$

In formula (3), C is the number of categories, $y_{i,j}$ are the actual labels, and $\hat{y}_{i,j}$ are the probabilities predicted by the model. In a distributed architecture, data is processed in a distributed manner through a specific framework, and multiple nodes are trained simultaneously to enhance work efficiency. This technology has been promoted in multiple industries such as medical imaging and intelligent driving, improving the accuracy and processing speed of classification [6].

3.3. Image Retrieval Based on Big Data Mining

Image retrieval aims to find results similar to query images from databases, and relying on big data mining Techniques, especially feature matching and the application of deep neural networks, greatly improve the accuracy and speed of retrieval. When conducting large-scale image searches, the widely used strategy is to evaluate similarity based on feature vectors. For any image, deep learning algorithms are used to extract its feature representation $V_i \in R^d$, and then similarity search is completed by comparing the spatial distances between these feature vectors. The commonly used distance calculation methods in this process are cosine similarity or Euclidean distance. For example, given the query image feature vector q and the database image feature vector vi, the Euclidean distance calculation formula is:

$$D(q,v_i) = \sqrt{\sum_{k=1}^{d}(q_k - v_{i,k})^2} \tag{4}$$

To enhance search efficiency, local sensitive hashing techniques can be used to achieve approximate search. In distributed computing systems, feature vectors are stored in data blocks, and parallel processing techniques are used to accelerate retrieval speed. The fusion clustering algorithm is used to classify feature vectors and further optimize the search performance of large-scale image databases. This strategy has been widely adopted by major e-commerce platforms and effectively applied in product image search and product recommendation systems. It also plays an important role in content search on social media platforms, enhancing user interaction experience and the system's ability to provide quick feedback.

4. Application of Big Data Analysis in Image Processing

4.1. Visualization Analysis of Image Data

Visualizing and analyzing image data plays a crucial role in the field of big data analysis. By presenting the basic attributes and unique features of images in a graphical manner, it assists researchers in understanding the distribution patterns and intrinsic properties of image data. The Table 1 listed below is a set of data examples that details the main feature data of five images, including the average brightness value, edge recognition score, and the main color composition of the images (red, green, blue).

Table 1. Image Data Analysis Table.

Image ID	Average brightness	Edge detection score	Main color component (red)	Main color component (green)	Main color component (blue)	Brightness-Edge Distance
Img001	128	0.80	125	120	115	12.00
Img002	145	0.75	140	135	130	5.00
Img003	132	0.85	130	128	125	8.00
Img004	155	0.90	155	150	145	15.00
Img005	140	0.78	135	132	128	0.04

The average brightness of an image refers to the arithmetic mean of the brightness values of all pixels. The evaluation value based on edge recognition technology can reflect the sharpness of the edges in the image. The main color composition of an image is to consider the average values of its red, green, and blue channels. When conducting statistical analysis, the Euclidean distance D was used to measure the difference between brightness and edge detection scores, and the formula is as follows:

$$D = \sqrt{(B - \bar{B})^2 + (E + \bar{E})^2} \tag{5}$$

By utilizing visual analysis of images, unique attributes and anomalies in image data can be quickly identified, laying a solid data foundation for subsequent image classification and search work. This technology is widely used for image quality assessment and in-depth analysis of image content.

4.2. Application of Big Data Analysis in Image Quality Assessment

When evaluating the effectiveness of image processing techniques, the detection of image quality is crucial, and big data analysis techniques play a decisive role in the comprehensive evaluation of such image quality. Taking the Table 2 below as an example, the quality rating elements of five images are detailed, covering peak signal-to-noise ratio (PSNR), image structure similarity index (SSIM), image compression degree, and image clarity index (Blurriness Index).

Table 2. Image Quality Evaluation Index Table.

Image ID	PSNR	SSIM	Compression ratio	fuzziness index	weighted quality score
Img001	35.5	0.85	2.5	0.45	14.745
Img002	40.2	0.92	2.8	0.38	16.690
Img003	38.8	0.89	2.7	0.42	16.104
Img004	42.1	0.95	3.0	0.35	17.485
Img005	37.4	0.87	2.6	0.41	15.527

To comprehensively evaluate image quality, the weighted quality score (WQS) formula is used:

$$WQS = w_1 \cdot PSNR + w_2 \cdot SSIM + w_3 \cdot Compression_{Ratio} - w_4 \cdot Blurriness_{Index} \tag{6}$$

In formula (6), $w_1 = 0.4$, $w_2 = 0.4$, $w_3 = 0.1$, and $w_4 = 0.1$ are weight factors. After analysis by the scoring system, the image with the number Img004 received the highest score (17.485 points), and its PSNR and SSIM index were both in the top positions, indicating that the image still maintains high-quality images and low blurriness even under

high compression ratios. With the help of big data analysis methods, the evaluation of image quality can be comprehensively analyzed from multiple perspectives and indices. This evaluation strategy plays a key role in many fields such as image compression, transmission, and enhancement processing.

4.3. Application of Data Analysis Techniques in Image Content Understanding

Analyzing images involves systematically constructing information about the elements, background, and underlying meanings within the painting. The use of big data parsing technology in this field has greatly promoted image parsing, with the core pursuit being to extract core attributes through comprehensive dimensional analysis to enhance recognition accuracy. Table 3 lists the analysis data of five images, including entity count, number of feature point locks, semantic score, and environment complexity level

Table 3. Index Table for Image Content Analysis.

Image ID	Number of objects	Key point detection count	Semantic score	Scene complexity	Content comprehension score
Img001	5	150	0.85	3	1.970
Img002	8	200	0.92	5	2.884
Img003	6	180	0.88	4	2.296
Img004	9	220	0.95	6	3.170
Img005	7	190	0.89	4	2.638

In order to comprehensively evaluate the performance of image content understanding, the Content Understanding Score (CUS) calculation formula is used:

$$CUS = w_1 \cdot OC + w_2 \cdot \frac{KP}{100} + w_3 \cdot SS - w_4 \cdot SC \quad (7)$$

In formula (7), $w_1 = 0.3$, $w_2 = 0.4$, $w_3 = 0.2$, and $w_4 = 0.1$ respectively represent the weights of object count (OC), keypoint detection count (KP), semantic score (SS), and scene complexity (SC). The formula is: $CUS = w_1 \cdot OC + w_2 \cdot KP + w_3 \cdot SS - w_4 \cdot SC$. The high weight indicates that identifying the number of objects and key points is essential to understanding the core content. For example, Img004 ranks among the top in both target counting and core element recognition indicators, with a score of 3.170. The semantic score maps the accuracy of extracting semantic information from images, and images with higher scores demonstrate excellent semantic parsing skills. Environmental complexity significantly influences the difficulty of content interpretation. Comprehensive analysis shows that big data technology has played a significant role in improving image content parsing capabilities. It has been widely applied in various fields such as autonomous driving, medical imaging diagnosis, and smart city construction, providing strong support for the efficient operation of intelligent systems [7].

4.4. Utilizing Big Data Analysis to Mine Semantic Information of Images

The mining of image semantic information aims to extract high-level semantics from visual data, revealing the essential content of images and their inherent logical connections. Big data analysis technology has demonstrated outstanding performance in object recognition, attribute extraction, and semantic association construction. The following Table 4 presents in detail the various indicators of semantic information parsing for five images, including the number of recognized objects, the number of main attributes, the score of semantic connections, and the strength of interference signals:

Table 4. Index Table for Image Semantic Information Analysis.

Image ID	Number of detected objects	Number of key features	Semantic relationship score	Noise level	Semantic information score
Img001	10	50	0.75	0.10	5.160

Img002	15	65	0.85	0.08	7.278
Img003	12	60	0.80	0.12	6.172
Img004	20	80	0.90	0.07	9.387
Img005	18	75	0.88	0.09	8.585

In order to comprehensively evaluate the mining effect of image semantic information, this paper adopts the Semantic Information Score (SIS) formula:

$$CUS = w_1 \cdot OC + w_2 \cdot \frac{KF}{100} + w_3 \cdot SS - w_4 \cdot SC \tag{8}$$

In formula (8), $w_1 = 0.3$, $w_2 = 0.4$, $w_3 = 0.2$, and $w_4 = 0.1$ respectively represent the weights of the number of detection objects (*DO*), the number of key features (*KF*), the semantic relationship score (*SRS*), and the noise level (*NL*). In the process of semantic information extraction, the number of targets and the number of feature points in the image play a core role. By utilizing big data analysis technology and integrating multidimensional indicator evaluation and advanced model construction, the efficiency of extracting semantic information from images can be enhanced. This technology has been widely applied in multiple fields such as text automatic generation, dynamic monitoring, and intelligent recommendation, providing strong technical support for intelligent scene analysis.

5. Conclusion

The image processing industry has gained strong impetus from the integration of big data mining and analysis technologies, which have achieved significant results in feature extraction, classification, search, and meaning analysis of images, providing innovative solutions to overcome the challenges faced by traditional image processing. Research has found that big data mining demonstrates efficient capabilities in accurately obtaining image features and improving processing speed. Big data analysis has outstanding advantages in intuitive display and deep semantic mining of image data. The results of this study provide a reference for the development of image processing technology, and also lay a solid foundation for the exploration and application of other data intensive fields.

References

1. H. M. Alzaabi, M. A. Alawadhi, and S. Z. Ahmad, "Examining the impact of cultural values on the adoption of big data analytics in healthcare organizations," *Digit. Policy Regul. Gov.*, vol. 25, no. 5, pp. 460–479, 2023, doi: 10.1108/DPRG-12-2022-0148.
2. A. R. Kulkarni, N. Kumar, and K. R. Rao, "Efficacy of bluetooth-based data collection for road traffic analysis and visualization using big data analytics," *Big Data Min. Anal.*, vol. 6, no. 2, pp. 139–153, 2023, doi: 10.26599/BDMA.2022.9020039.
3. M. E. Fernandez, et al., "How is Big Data reshaping preclinical aging research?," *Lab Anim.*, vol. 52, no. 12, pp. 289–314, 2023, doi: 10.1038/s41684-023-01286-y.
4. X. Lai, et al., "Efficiency scoring for subway tunnel construction based on shield-focused big data and Gaussian broad learning system," *J. Constr. Eng. Manag.*, vol. 149, no. 12, Art. no. 04023132, 2023, doi: 10.1061/JCEMD4.COENG-13170.
5. J. Wang, "Cost control problems and countermeasures of e-commerce enterprises under the background of big data and Internet of Things," *J. Comput. Methods Sci. Eng.*, vol. 23, no. 6, pp. 3135–3145, 2023, doi: 10.3233/JCM-226931.
6. M. Park and N. P. Singh, "Predicting supply chain risks through big data analytics: role of risk alert tool in mitigating business disruption," *Benchmarking Int. J.*, vol. 30, no. 5, pp. 1457–1484, 2023, doi: 10.1108/BIJ-03-2022-0169.
7. N. Li, et al., "Modeling categorized truck arrivals at ports: Big data for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 2772–2788, 2022, doi: 10.1109/TITS.2022.3219882.