Significance and Challenges of Image Mining: A Comprehensive Study

Shilpa. K
Assistant Professor, Department of Computer Science, Government Science College, N.T.Road, Bengaluru, India

Abstract- Image mining is an interdisciplinary field encompassing machine vision, image processing, image retrieval, data mining, machine learning, databases, and artificial intelligence. With advancements in digital technology, the volume of data is continuously increasing. A major challenge is creating dedicated databases for images, which is complex and demands significant attention. Fields such as medicine, geographical systems, robotics, health sciences, and engineering require separate databases for image storage, highlighting the crucial role of image mining in research and development. The core idea of image mining is the extraction and discovery of new information or knowledge from databases containing images. Unlike general data mining, image mining focuses specifically on extracting information from images and identifying relationships between image sets and other patterns based on user requirements. While various algorithms have been developed for image mining, further work is needed to enhance the specificity, precision, accuracy, and effectiveness of the results. This paper examines the image mining process, its applications, and the challenges it faces, while also exploring future research directions in the field.

Keywords— Data Mining, Image Mining, Knowledge Discovery, Machine Learning, Artificial Intelligence, Rule Mining, Datasets.

1. INTRODUCTION

Image mining, an extended version of data mining within the image domain, integrates expertise from computer vision, image processing, image acquisition, image retrieval, data mining, machine learning, databases, and artificial intelligence (AI). Every day, an immense volume of images is captured or generated, particularly in fields such as medicine—where MRI scans, CT scans, ultrasound scans, mammogram images, and X-ray images are prevalent. In meteorology, the challenges intensify as satellite images must be stored and processed for accurate weather forecasting. This shift has significantly impacted databases, raising challenges related to image storage, indexing, querying, and the formulation of queries for effective knowledge extraction. Historically, data was confined to alphanumeric characters, but now it encompasses images, redefining the data landscape.

Image mining focuses on extracting patterns from vast image collections, distinguishing it from image processing, which emphasizes specific characteristics of individual images. The rapid production and storage of high volumes of images, including satellite images, medical images, and digital photos, in cloud environments necessitate efficient analysis to glean valuable insights. The fundamental challenge lies in analyzing the raw pixels in images or sequences of images to detect objects and their interrelationships.

The primary objective of image analysis is to uncover significant patterns in images without needing detailed knowledge of their content. This means that valuable patterns can be extracted from a series of images as input, even without a deep understanding of the images' specific content. The advancement in digital technologies has further accentuated the importance of image mining, making it a critical area of research and development across various domains.

2. CONTENT-BASED IMAGE RETRIEVAL (CBIR)

Image mining can be performed manually by segmenting and analyzing data to identify specific patterns or through automated programs that analyze the data. In context-based image retrieval systems, primary descriptors such as color, texture, and existing shapes are crucial for identifying and retrieving similar

images from large image databases, making manual extraction impractical due to the sheer volume of data. Content-Based Image Retrieval (CBIR), also known as Query by Image Content (QBIC) and content-based visual information retrieval (CBVIR), uses machine vision to retrieve digital images from extensive databases. Traditional image retrieval methods, like indexing, are time-consuming and inefficient, relying on keywords or numerical identifiers linked to classified descriptions. These older methods do not leverage CBIR content.

In CBIR, each image stored in the database has unique characteristics that are extracted and compared with the features of the query image. This approach combines knowledge from various fields, including pattern recognition, object matching, machine learning, and microwave filtering. CBIR aims to discover visual properties of images without relying on descriptive text.

CBIR systems compare database images to a query image, focusing on developing techniques that enhance digital libraries of images based on features automatically extracted from queries. These features can be categorized as low-level (e.g., color, texture, edge, shape) or high-level characteristics. Unlike text-based image retrieval systems (TBIR), which index and retrieve images based on descriptions like size, type, date, time of capture, owner, keywords, or other explanatory texts, CBIR uses visual attributes for retrieval.

In a typical CBIR system, as illustrated in Figure 1, visual concepts are extracted from databases, and features are represented as multi-dimensional vectors. To retrieve an image, users provide a sample image as input, which the system converts into an internal feature vector. The similarity between the input image and database images is calculated, facilitating search and indexing, and ultimately enabling efficient image retrieval based on patterns.

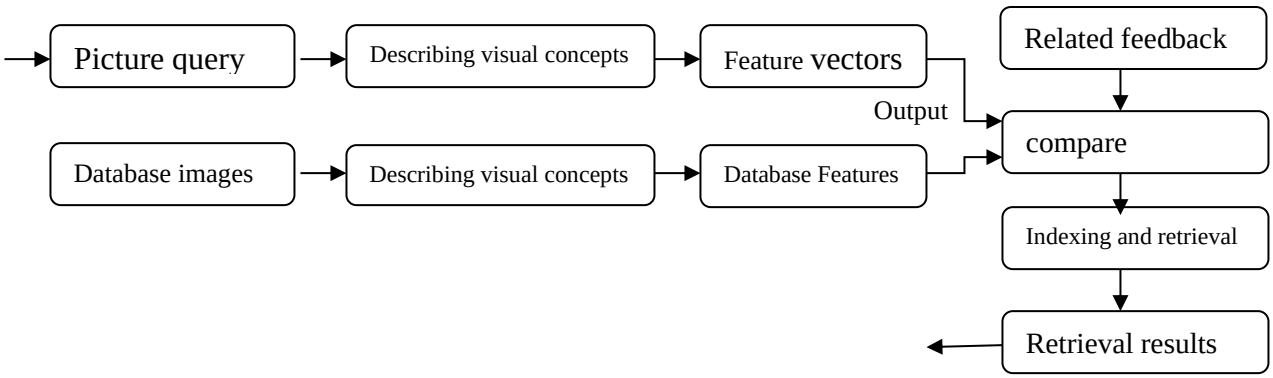

Figure 1. An example system architecture Content-Based Image Retrieval CBIR.

3. IMAGE MINING

Image mining involves the extraction of implicit knowledge, relationships between image data, or other patterns that are not explicitly stored within the images. Unlike other image processing techniques, image mining does not aim to detect specific patterns in images. Instead, it focuses on identifying and uncovering patterns and deriving knowledge from images within a set, based on low-level (pixel) information. As a research field, image mining has evolved into an interdisciplinary area, combining the expertise and tools of data mining, databases, computer vision, image processing, image retrieval, statistics, pattern recognition, machine learning, and artificial intelligence.

Figure 1 below gives a glimpse of image mining process.

The image mining process consists of several components, including:

- Image analysis, which covers image pre-processing, object recognition, and feature extraction.
- Image classification.
- Image indexing.
- Image retrieval.
- Data management.

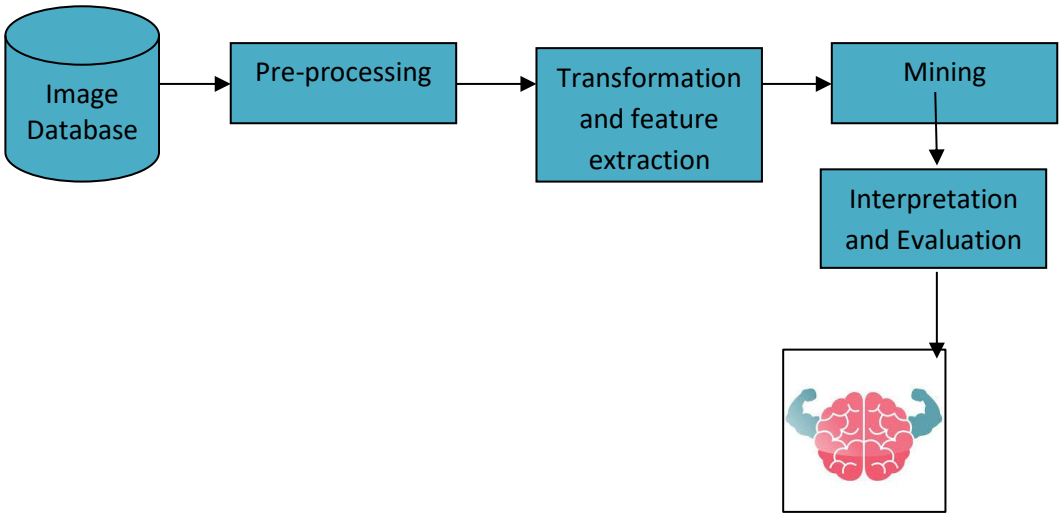

Figure 1: Image mining process.

3.1 Image pre-processing

The initial step in image mining is image pre-processing, which operates at the lowest level of abstraction. During this stage, the image is processed to remove distortions and enhance its resolution for further analysis. For instance, in satellite imagery, pre-processing is used to extract the spatial locations of fire spots, represented by latitude and longitude (Figure 2). Various pre-processing algorithms, such as noise reduction, averaging, median filtering, and Wiener filtering, are employed to minimize noise and improve image quality.

Figure 2: Natural-colour image of the Thomas Fire in Ventura County California. Photo was taken by NASA's Aqua satellite. Credit: NASA Goddard LANCE/EOSDIS MODIS Rapid Response Team

3.2. Classification

Classification is a supervised method of data grouping. In supervised methods, a learning set of labeled images is provided for classification. This process typically involves two phases: the learning phase and the testing phase. In the learning phase, images are distinctly profiled, and learning is based on their classes. In the testing phase, parts of the specifications are used to classify images. The most popular classification methods include decision trees, Bayesian classifiers, SVM-based classification rules, neural networks, and fuzzy logic techniques.

One crucial method in the classification process is the use of decision trees. Decision trees divide the decision space into smaller areas based on the entire sample, effectively breaking down complex decisions into simpler, uniform results. This approach naturally reflects the recognition strategies used in human decision-making processes.

3.3. Color Processing

One method of color image processing involves using a color histogram. A color histogram represents the color distribution within an image and can be applied to the entire picture or specific ranges within it. For an RGB color image, which is an M * N * 3 array of color pixels, each color pixel is a triplet specifying the amounts of red, green, and blue in the image. This can be visualized as a stack of three black and white images, where the entries for red, green, and blue are combined to form a color image. The average of each color component in the image can be calculated (Formula 1).Average pixels red = R (P) / P

Average green pixels = (G (P)) / P

Average blue pixels = (B (P)) / P

Formula1: Calculation formula

Where P is the total number of image pixels. R (P) is the number of red pixels. G (P) is the number of green pixels and B (P) is number of blue pixels.

3.4. Clustering

Clustering, a branch of machine learning, is an unsupervised method that involves automatically dividing samples into groups, or clusters, whose members are similar to each other. A cluster is a collection of objects that are similar to each other, while objects in different clusters are dissimilar. Various criteria are considered in clustering, such as the proximity of objects, which leads to distance-based clustering.

Clustering divides a heterogeneous population into a number of homogeneous subsets or clusters. Unlike categorization, clustering does not rely on pre-determined categories. In model-based categorization, data is allocated to pre-defined categories determined by prior research (e.g., gender, skin color). In contrast, clustering groups data based on similarity without predefined categories, and the user determines the titles of each group. For instance, clusters of symptoms may indicate different diseases, and clusters of customer features may reveal distinct market segments. Clustering is often used as a precursor to other data mining analyses or modeling.

3.5 Feature Extraction

Image mining involves compressing extracted information from known objects into attributes. Both global and local descriptors are used to represent images. Global descriptors are easier to compute and result in minimal segmentation errors, while local descriptors provide precise information and help identify distinct patterns. These features are typically represented numerically, offering a complex mathematical representation of an image's shape, texture, color, and more. According to Martinez, descriptors can be categorized into the following description tools:

a. Basic elements: Tools used by relevant descriptors include grid layout, time series, 2D/3D multiple views, spatial 2D/3D coordinates, and temporal interpolation.

b. Color: The color histogram (Figure 4) is a graphical representation of the color distribution within an image, offering effective characteristics for easy computation. As a descriptor, the histogram cannot be rotated or translated. Color descriptors identified by Martinez include color space, color quantization, scalable color, dominant color, color layout, color structure, and group-of-frames/group-of-pictures color.

3.6 Selecting Properties

When selecting properties, methods based on entropy, Gain ratio, Gini index, chi-square, among others, are utilized. These methods help discretize properties using techniques such as chi-merge discrete cut points, MDLP, or LVQ. When employing decision trees for classification, these discretization methods establish

one or more intervals, influencing the creation of binary or multi-class decision trees that are both accurate and compact. Evaluation techniques like n-fold cross-validation or train-test methods are commonly used to assess these trees.

Feature selection plays a crucial role in reducing problem dimensions, enhancing prediction accuracy, and reducing computational time. This involves removing irrelevant, redundant, and noisy features, aiming to select a subset of features. Typically, feature selection employs various search algorithms, including sequential forward selection, sequential backward selection, genetic algorithms, particle swarm optimization, and branch and bound feature optimization, which are widely utilized for this purpose.

3.7. Histogram Equalization

Histogram equalization is a technique used in image processing to adjust contrast. It redistributes the intensity values in the image's histogram, thereby enhancing contrast more uniformly. This process improves the contrast in regions with initially lower local contrast, making it particularly beneficial for images like radiology scans where distinct foreground and background are crucial.

Another histogram method in image processing is severity histogram. This method considers features such as average, variance, skewness, elongation, entropy, and energy to characterize the distribution of intensity or other image attributes.

Figure 3: Simple histogram of flowers.
Photo courtesy: Sanjaya. A
Histogram courtesy: PineTools

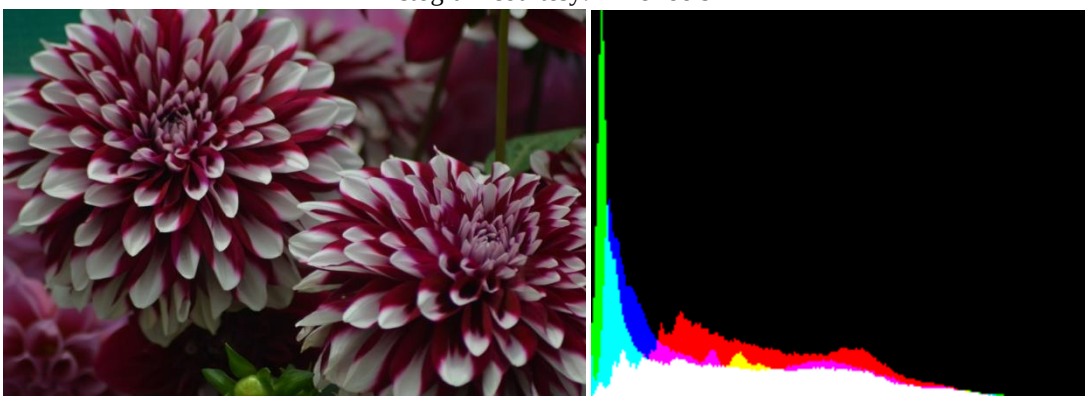

Figure 4: Colour histogram of flowers.
Photo courtesy: Sanjaya. A,
Colour histogram courtesy: Lunapic

4. Automating image analysis and subsequent knowledge acquisition through computer-driven image processing holds immense potential, yet the field is still in its early stages with numerous areas requiring further exploration. Several critical issues must be addressed to enable efficient image analysis and knowledge derivation:

- Moving beyond the low-level pixel representation of images to develop representations capable of

encoding the latent information embedded within an image is crucial.

- Classifying the patterns extracted in Image Mining remains a challenge, particularly automating the derivation of appropriate decision criteria for clustering image representations.
- Implementing suitable indexing methods and establishing standards for indexing procedures to efficiently retrieve knowledge from images are essential concerns.
- Developing a query language capable of handling both visual patterns and textual information is necessary.
- Analyzing and retrieving knowledge from images stored online presents a significant challenge for image mining.

5. CONCLUSION

Every day, sources like satellite imagery, medical scans, and digital images generate vast amounts of valuable information. The sheer volume and complexity of these data make manual analysis for extracting useful patterns and making informed decisions nearly impossible. Image mining emerges as a promising field for extracting knowledge from images, yet it remains in its nascent stages. Further research is essential to advance techniques in image processing, feature extraction, image segmentation, and object identification.

This paper has outlined the unique aspects of image mining, discussed the general process of image analysis, and explored key techniques in the field. Additionally, we introduced image mining as a cutting-edge research area within imaging databases. Various methods and techniques proposed by researchers for image mining have also been reviewed. Moving forward, continued exploration and development in image mining will be crucial for harnessing the full potential of visual data across diverse applications.

REFERENCE

[1] Fayyad U, Shapiro G, Smyth P. Knowledge Discovery and Data Mining [Online]. 2011 [Cited2011Aug8]; Available from: URl: http://www. Aaai.org.

[2] Tan J. Medical Informatics: Concepts, Methodologies, Tools, and Applications. Hershey: IGI Global snippet; 2008.

[3] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references).

[4] LaTour KM, Eichenwald S. Health Information Management: Concepts, Principles, and Practice. Chicago: AHIMA; 2002. p. 478-80.

[5] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Philadelphia: Elsevier; 2011.

[6] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[7] Chen H, Fuller SS, Friedman C, Hersh W. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. New York: Springer; 2005.

[8] Maimon OZ, Rokach L. Data Mining And Knowledge Discovery Handbook. New York: Springer Science & Business; 2010. p. 1.

[9] Chakrabarti S, Cox E. Data Mining: Know It All. Amsterdam: Morgan Kaufmann p. 7; 2009.

[10] J. Zhang, W. Hsu, M. Lee, Image Mining: Issues, Frameworks And Techniques, In Proc. Of the second International workshop on Multimedia Data Mining, SanFrancisco, USA, August 2001.

[11] C. Ordonez, E. Omiecinski, Image Mining: A new approach for data mining", Technical Report GIT-CC-98-12, Georgia Institute of Technology, College of Computer, 1998.

[12] Ji Zhang, Wynne Hsu, Mong Li Lee, "An Information. Driven Framework For Image Mining", Computer Science, School of Computer, National University of Singapore, IEEE, August 2001.

[13] Ramadass Sudhir, "A Survey on Image Mining Techniques: Theory and Applications", Computer Engineering and Intelligent Systems, Vol2,No,6,2011.

[14] Monika Sahu, Madhu P Shrivastava, Dr. M A Rizvi, "image mining: a new approach for data mining based on texture",IEEE,2012.

[15] A.Kannan, DR.V.Mohan, Dr.N.Anbazhagan, "Image Clustering and Retrieval using Image Mining Techniques",IEEE,2010.

[16] Nishchol mishra1,Dr.sanjay Silakari, "Image Mining the Context of content based Image Retrieval:A perspective",IJCSI,Vol.9,Issue4, No3, July 2012.
[17] Sanjay T.Gandhe, K.T. Talele, and Avinash G.Keskar. "Image Mining Using Wavelet Transform". Springer-Verlag Berlin Heidelberg 2007.
[18] Tomas Berlage, "Analyzing and mining image database", DRUG DISCOVERY TODAY: BIOSILICO, DDT, Vol 10, Number 11, June 2005.
[19] Petra Perner, "Image mining: issue, framework, a generic tool and its application to medical-image diagnosis", Elsevier,2002.
[20] Aswini Kumar Mohanty, Manas Ranjan Senapati, Saroj Kumar Lenka, " A novel image mining technique for classification of mammograms using hybrid feature selection, "Springer, 23 February 2012.
[21] Chidansh Amitkumar Bhatt, Mohan S.Kankanhalli, "Multimedia data mining: state of the art and challenges", springer Science +Business Media, LLC 2010.
[22] A.Hema,E.Annasaro,"a survey in need of image mining techniques", International Journal of Advanced Research in Computer and Communication Engineering Vol.2,Issue2, February 2013.