Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/38/20230532

Research on unsupervised image retrieval methods based on contrastive learning

Hanhong Liu

School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong, China

hanhong.liu@connect.polyu.hk

Abstract. In the convergence of fashion and artificial intelligence (AI), significant strides have been made in areas such as clothing recognition, retrieval, and classification, enabled by advanced AI technologies and expansive annotated datasets. As the AI in Fashion market continues to surge, the future of the fashion industry promises to be redefined by intelligent, efficient, and more accessible solutions. Image retrieval, one of the important parts in AI, has experienced remarkable growth, empowered by advanced algorithms and vast annotated datasets, making it a crucial component in various domains such as digital libraries, online marketing. Therefore, this report mainly provides an extensive review of image retrieval methods and the emerging paradigm of contrastive learning, underscoring their relevance and applications in the realm of artificial intelligence. This paper primarily reviews the technologies in the amalgamation of the image retrieval field and contrastive learning. It elucidates the history and progression of image retrieval, offers a methodical analysis of the two primary approaches—text-based image retrieval and content-based image retrieval—and examines how contrastive learning is employed in image retrieval systems.

Keywords: Unsupervised Learning, Image Retrieval, Contrastive Learning.

1. Introduction

With the rapid development of digital technologies and the camera industry, the digital image database is increasing, and the digital database management has gained a lot of attention, including medical, artwork, fashion, and so on. Based on the above requirement, image retrieval is developing, a key area of research in computer vision that aims to develop algorithms that can automatically search and retrieve images from large databases (Li et al., 2021) [1].

Nowadays, due to the more complex operation and implementation process, content-based image retrieval is the main research area to improve its accuracy and user experience. The methods from traditional methods extracting features using HOG, Haar, LBP features and train the model with adaboost algorithm, random forest algorithm and so on, then, the model will be trained to do the classification and retrieval. In 2015, deep learning became the most popular method in many areas and in computer vision, the method shifted from SIFT to CNN as well. Among the new techniques that have emerged, contrastive learning is particularly promising. By learning robust and discriminative feature representations through comparing similar and dissimilar pairs, contrastive learning has demonstrated excellent performance in several tasks, including unsupervised and self-supervised

DOI: 10.54254/2755-2721/38/20230532

learning scenarios. This report investigates these methods in detail, emphasizing their underpinning mechanisms, the challenges they face, and the cutting-edge solutions that address these issues.

Therefore, this report mainly reviews image retrieval methods and the emerging paradigm of contrastive learning, underscoring their relevance and applications in artificial intelligence.

2. Image Retrieval

Regarding the developmental progression of image retrieval technology, image retrieval technology can be classified into two primary stages: the text-based image retrieval stage and the content-based image retrieval stage. Figure 1 depicts the flow charts for two typical methods of image retrieval processing. The system of text-based image retrieval (TBIR) utilizes images from the database that are already enriched with annotations, keywords, or descriptions. These elements are then employed to correlate with the textual input provided by users, subsequently generating relevant concepts for user output. The content-based image retrieval (CBIR) system is divided into two stages: offline stage feature extraction and online stage feature extraction. As shown in figure 2, the CBIR system automatically extracts and stores the features of each image in the database as feature vectors during offline stage and extracts and represents the features of the query image during online stage when there are input images by users. Then, the system assesses the likeness between the feature vectors of the images stored in the database and those of the query image. Subsequently, an indexing scheme is applied to facilitate an efficient image database search, and the retrieval process is performed accordingly. Finally, the system returns the images that resemble the query image most [2].

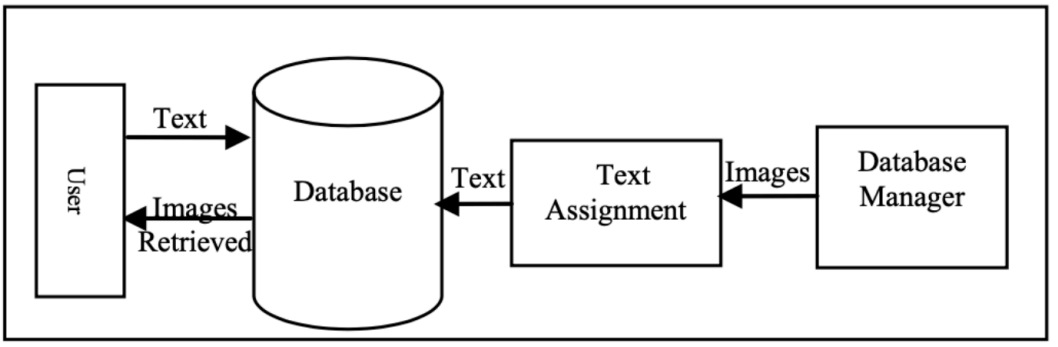

Figure 1. A typical TBIR system [2].

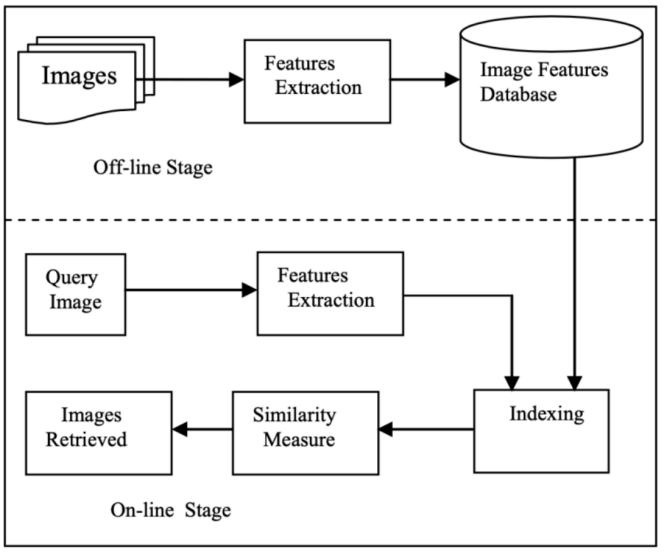

Figure 2. A typical CBIR system [2].

DOI: 10.54254/2755-2721/38/20230532

2.1. Text-based Image Retrieval

The development of TBIR dates back to the late 1970s. The common approach to image retrieval often involved adding textual annotations to images and subsequently employing text-based database management systems (DBMS) to facilitate the retrieval procedure [3]. Query Expansion, Relevance Feedback and Latent Semantic Analysis are the representatives of TBIR and are widely used in the industry.

In text-based image retrieval, the system retrieves images based on the text query entered by the user. TBIR involves manually annotating images in a database with descriptive information such as annotations, keywords, or descriptions. This annotation process captures the image content and metadata, including the image filename, format, size, and aspects. By providing such textual information, TBIR enables more efficient searching and retrieval of images based on their content and associated attributes. [2] Finally, the system searches for images with the same or similar keywords as the query and output it. The text-based method was mentioned earlier which exists numerous disadvantages: 1) manually annotating huge databases is not practical; 2) the end user must add annotations, exposing this method to human perception; and 3) these annotations are applicable to only one language [4]. One of the earliest approaches to text-based image retrieval used manually assigned textual descriptions of images [5]. This approach is limited by the subjective nature of textual descriptions, which are often incomplete and inconsistent.

To overcome these limitations, the researchers developed automatic methods for generating textual descriptions from images. The other main methodology for text-based image retrieval is to use image captions, which are short textual descriptions of images often used in social media and photo-sharing platforms. Image captions can be used efficiently for text-based image retrieval, according to recent research. An approach for cross-modal image retrieval, for instance, that combines visual data and textual features taken from image captions was proposed by Li et al. [6]. The method achieved cutting-edge outcomes on various benchmark datasets.

The deep learning method is widely used in text-based image retrieval through the development of AI technologies and the application of neuron networks. The computer system learns to place similar images and texts next to one another in a common embedding area. Numerous research have shown that this strategy works.

2.2. Content-based Image Retrieval

The other method is content-based image retrieval (CBIR). This approach has been strongly advocated as a potent solution to the limitations inherent in text-based image retrieval methods. Text-based image retrieval primarily relies on accurate, detailed, and exhaustive annotations to function effectively, a requirement that often proves to be unrealistic and resource-intensive. This predicament underscores the necessity for an alternative that can function independently of such exhaustive text annotations - a role well-suited for CBIR.

Content-Based Image Retrieval (CBIR) is crafted to examine the tangible content within the image, encompassing aspects like colour, texture, shape, and other intrinsic information extracted directly from the image. It harnesses the power of sophisticated image processing and machine learning algorithms to understand and learn from these features. The benefits of CBIR go beyond merely circumventing the limitations of text-based methods. By directly analysing the image content, CBIR allows for a richer, more nuanced retrieval process that can handle a variety of complex and subtle image queries, thereby significantly expanding the scope and capabilities of image retrieval technologies.

CBIR research began in earnest in the early 1990s, with researchers indexing images based on visual features such as texture and color. Several algorithms and image retrieval systems were proposed during this period. One straightforward strategy is to extract the global descriptor of an image. This approach received significant attention and research emphasis within the image retrieval community during the 1990s and the initial years of the 2000s. However, the global descriptor approach is notoriously difficult to achieve in lighting, deformation, occlusion, and cropping situations.

DOI: 10.54254/2755-2721/38/20230532

These defects also lead to the low accuracy of image retrieval and limit the application scope of global descriptor algorithm. Just at this time, image retrieval algorithm based on local features brings the dawn to solve this problem [7]. The figure 3 show the milestone moments in case retrieval tasks over the years, and the proposed time based on SIFT feature and CNN feature algorithm is highlighted in the figure 4. The year 2000 is the end of the most physical traditional methods and then in 2003, it comes to the new stage which a lot of tasks released by computational methods and Artificial Intelligence. The traditional image retrieval methods, SIFT-based feature extraction applied to the image classification and started from 2012, the CNN-based method appeared with Alex Krizhevsky and his team utilized a neural network model known as AlexNet to set a new standard in image recognition, which can be recognized as a significant achievement for the field of computer vision. Their efforts achieved the highest global recognition accuracy recorded during the ImageNet Large Scale Visual Recognition Challenge (ILSRVC) in 2012. Convolutional neural networks, known for their efficiency in processing grid-like data, have been widely recognized for their applicability in image and video recognition tasks. Their unique architecture, designed to mimic the human visual system, has made them an area of intense study and application in the years following the 2012 achievement of AlexNet [7].

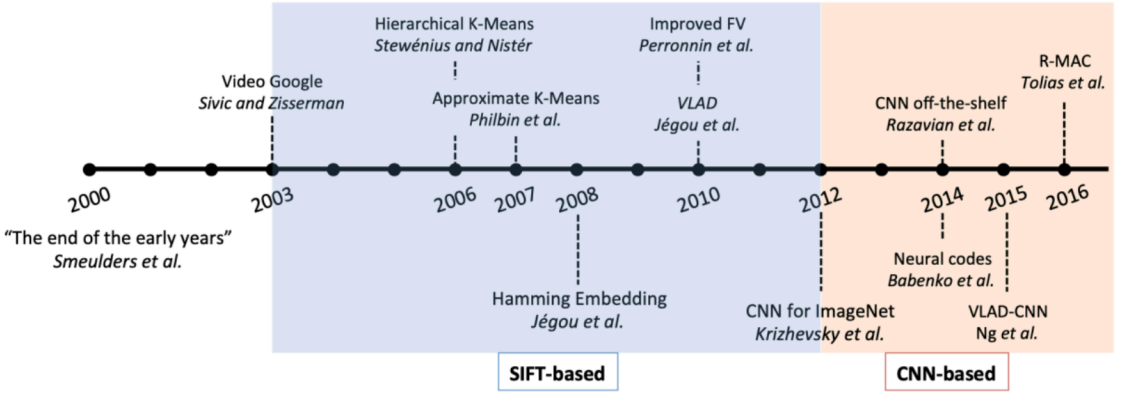

Figure 3. Milestones of instance retrieval [7].

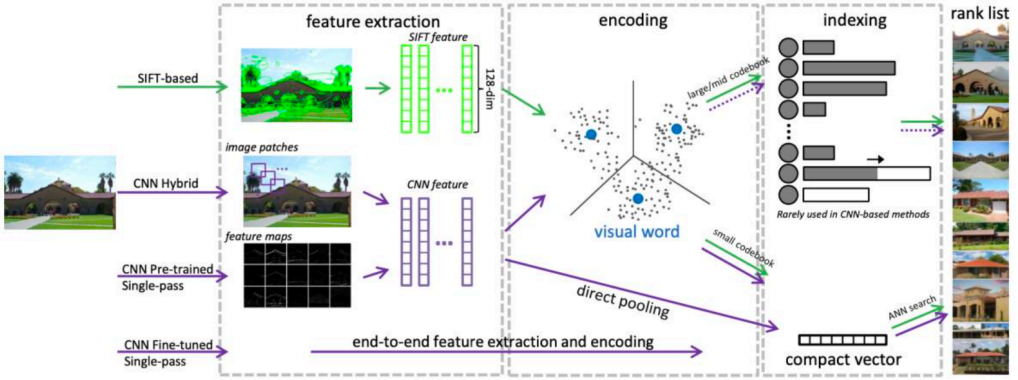

Figure 4. A general pipeline of SIFT-based and CNN-based retrieval models [7].

3. Contrastive Learning

A Contrastive learning has emerged as a popular approach in the field of unsupervised learning and Contrastive learning is a powerful and potential learning paradigm in deep learning. The figure 5 shows the detailed classification of deep learning, which can be modified into two areas: supervised learning and unsupervised learning. Contrastive learning is a branch of deep learning. The other popular research area of unsupervised learning is generative learning using GAN and VAE. The core

DOI: 10.54254/2755-2721/38/20230532

idea of this technique is to efficiently learn meaningful feature representations by training a model to distinguish (or compare) similar and dissimilar data samples. Contrastive learning aims to learn a representation that maximizes the similarity between positive and negative sample pairs while minimizing the similarity between negative and positive sample pairs. Although the basic concept of contrastive learning has existed for decades, its application in deep learning is not so far away. This method proves efficacious in diverse applications, spanning natural language processing, speech recognition, and image classification [8].

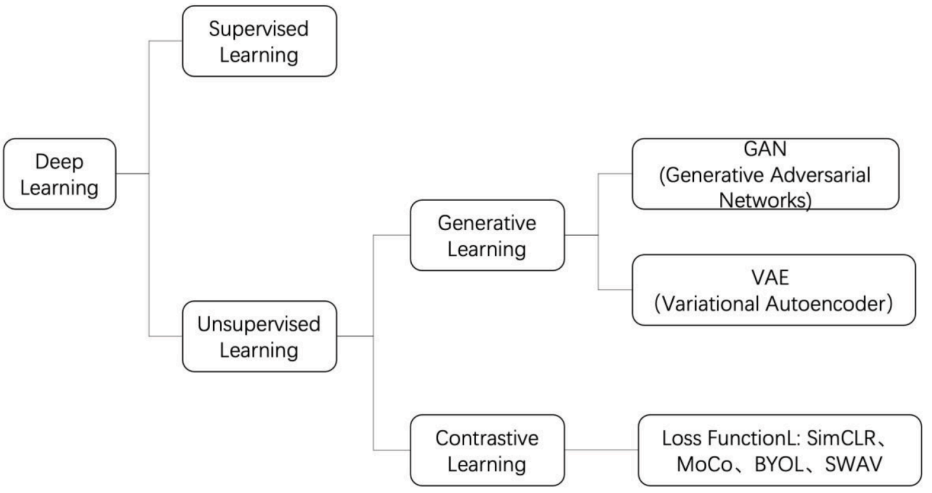

Figure 5. Classification of Contrastive Learning.

The fundamental concept behind contrastive learning revolves around the notion of grouping patterns. As shown in figure 6, the objective of contrastive learning entails training an encoder to represent similar data with a high degree of similarity, while distinguishing data from different classes with maximum dissimilarity. Therefore, contrastive learning is training the model to reduce the distance from the positive sample and expand the distance from negative sample. The training goal is to achieve the original space between the distribution's true distance by making the distance of the positive sample and the anchor point is significantly less than the distance of the negative sample and the anchor point, or by significantly increasing the similarity between the positive sample and the anchor point. The formula below shows such idea, which d() represents the distance:

$$d\left(f(x), f(x^{+})\right) \ll d\left(f(x), f(x^{-})\right) \tag{1}$$

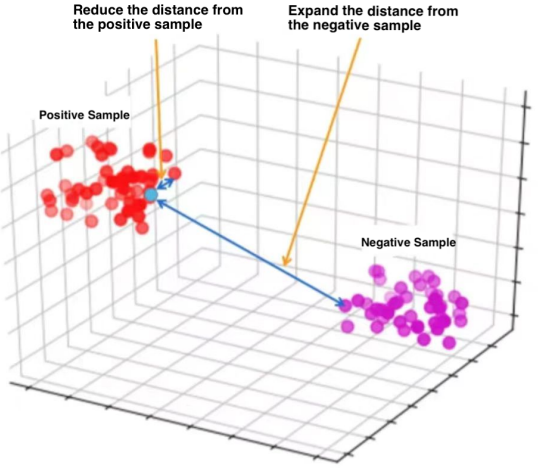

Figure 6. An introduction to contrastive visual representations.

DOI: 10.54254/2755-2721/38/20230532

Chen et al. introduced the "SimCLR" technique through their formulation of a straightforward framework for contrastive learning of visual representations. It is a typical negative example contrastive learning method to make the contrastive learning widely used in image recognition, image retrieval and other tasks. SimCLR enhances the learning of representations by optimizing the concurrence between distinct augmentations of the same data example, employing a contrastive loss function in the latent space, as shown in Figure 7 [9].

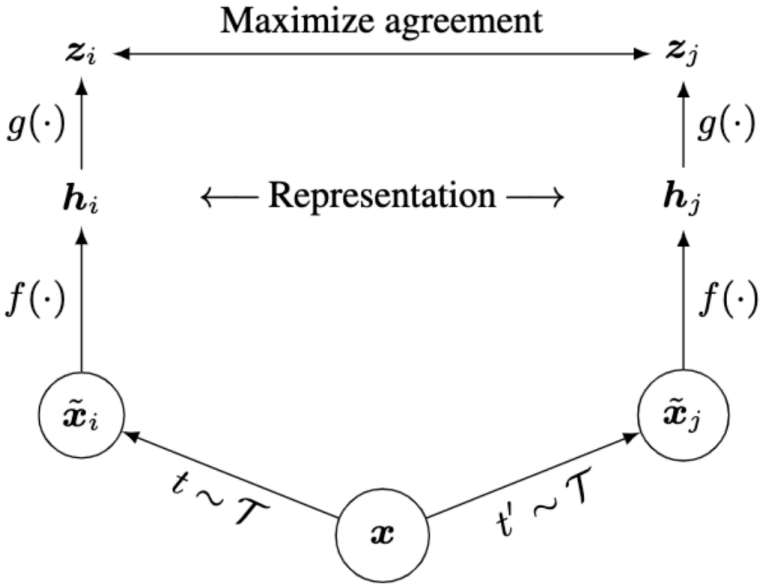

Figure 7. A basic framework for learning contrastive visual representations [9].

They experimented using dog image and experimented follow below steps:

There are two distinct operators sampled from identical set of augmentations ($t \sim \mathcal{T}$ *and* $t' \sim \mathcal{T}$) and each data example undergoes these operations to yield two interconnected perspectives.

1. Randomly enhance a dog image to generate two images.
2. These two enhanced images are sent to the network (·), producing two eigenvectors h.
3. h passes through the MLP (which is an fully connected network),that is, the projection operation $g(\cdot)$, and then generates z.
4. Then, use z to compute the contrastive loss.
5. After training, projection head (·) and use encoder (·) will be thrown. Away and representation h will be used for downstream tasks.

After processing the date through the contrastive learning framework, we will let sim(u, v) = u > v/ ‖ u ‖ ‖ v ‖ represent dot product l . Therefore, the loss function for a positive pair of samples (i,j) can be represented as below formula[10]:

$$l_{i,j} = -log \frac{\exp\left(\frac{sim(z_i,z_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{I}_{[k\neq i]} \exp\left(\frac{sim(z_i,z_k)}{\tau}\right)} \tag{2}$$

4. Conclusion

Image retrieval technology has evolved significantly over the years, and it can be classified into two primary stages: TBIR and CBIR. In image retrieval, the application of contrastive learning, with the process of the image augmentation, learns more discriminative feature representations. TBIR relies on manually annotated text information associated with images to facilitate searching and retrieval. However, it has limitations, such as the need for extensive manual annotations and language

DOI: 10.54254/2755-2721/38/20230532

dependency. On the other hand, CBIR overcomes these limitations by analyzing the actual content of images, such as color, texture, and shape, using advanced image processing and machine learning algorithms. Throughout the history of image retrieval, researchers have made notable advancements in both TBIR and CBIR. TBIR has seen improvements in query expansion, relevance feedback, and the use of image captions for more effective retrieval. Conversely, CBIR has witnessed significant progress with the introduction of local feature-based approaches, such as SIFT (Scale-Invariant Feature Transform), and the revolutionary impact of convolutional neural networks (CNN) in image recognition tasks. Integrating deep learning methods, especially CNNs, has further enhanced the performance of CBIR, allowing the system to learn to map images and text into shared embedding spaces, leading to more accurate and efficient retrieval results. As AI and computer vision advances, we can expect image retrieval technology to become even more sophisticated and powerful. Future research may explore hybrid approaches that combine the strengths of TBIR and CBIR or leverage other emerging technologies to address the remaining challenges in image retrieval, making it an indispensable tool for various applications across different domains.

References

[1] Li X, Yang J and Ma J 2021 *Neurocomputing* **452** 675–89
[2] Alkhawlani M, Elmogy M and El-Bakry H 2015 *International Journal of Computer and Information Technology* **4** 58–66
[3] Rui Y, Huang T S and Chang S-F 1999 *Journal of Visual Communication and Image Representation* **10** 39–62
[4] Liu Y, Zhang D, Lu G and Ma W-Y 2007 *Pattern Recognition* **40** 262–82
[5] Unar S, Wang X, Zhang C and Wang C 2019 *IET Image Processing* **13** 515–21
[6] Li X, Yang J and Ma J 2021 *Neurocomputing* **452** 675–89
[7] Zheng S, Yang F, Kiapour M H and Piramuthu R 2018
[8] He K, Fan H, Wu Y, Xie S and Girshick R 2020 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA: IEEE) 9726–35
[9] Chen T, Kornblith S, Norouzi M and Hinton G E 2020 *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020*, Virtual Event Proceedings of Machine Learning Research vol **119** 1597–607
[10] Bastanlar Y and Orhan S 2022 *Artificial Intelligence* vol **12**