Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 2, pp. 403–413
DOI : 10.47974/JIOS-1923

Comprehensive review and analysis on multi modal image retrieval

Pilli Mounika *
Department of Computer Science and Engineering
Jawaharlal Nehru Technological University
Kakinada 533003
Andhra Pradesh
India

K. Venkata Subba Reddy †
Department of Computer Science & Engineering (AI&ML)
Vidya Jyothi Institute of Technology
Jawaharlal Nehru Technological University
Hyderabad 500075
Telangana
India

N. Ramakrishnaiah §
Department of Computer Science and Engineering
University College of Engineering
Jawaharlal Nehru Technological University
Kakinada 533003
Andhra Pradesh
India

Abstract

Multimodal image retrieval, which involves retrieving images using various modalities such as text, audio, or other images, has significant research importance due to its wide-ranging applications and potential to enhance user experiences across multiple domains. Traditional image retrieval systems, content-based image retrieval (CBIR) systems rely solely on visual features, which can be limiting. By integrating multiple modalities, multimodal retrieval systems can offer more robust and accurate results. A research problem is considered multimodal when it integrates information from more than one type of data source. In multimodal image retrieval (MMIR) systems, one form of data is used to search

* *E-mail:* mounika260328@gmail.com (Corresponding Author)
† *E-mail:* kvsreddy2012@gmail.com
§ *E-mail:* nrkrishna27@gmail.com

for outcomes in the same or different modalities. In contrast, cross-modal systems strictly retrieve information from a different modality. A significant challenge in these systems is the effective comparison of input-output queries from different data types, due to their basic forms and the subjective nature of content similarity. Researchers have proposed various techniques to address this challenge and to bridge the semantic gap in information retrieval across different modalities. This article presents comprehensive analysis of various research works pertained in this multimodal image retrieval. The results and comparative analysis of current research works on benchmark datasets have also been discussed. At the end of the paper, some of open issues are presented for future research directions.

Subject Classification: *68P30.*

Keywords: *Content based image retrieval (CBIR), Multimodal, Cross-modal, Deep model, Multimodal image retrieval (MMIR).*

1. Introduction

Multimodal image retrieval focuses on retrieving the more similar images from multimedia upon considering the queries for retrieval from different modalities like text, query image, speech. The integration of multiple modalities, particularly visual and textual data, offers a promising avenue to enhance image retrieval systems. Multi-modal image retrieval leverages the complementary strengths of visual features (e.g., color, texture, shape) and textual descriptions (e.g., captions, tags, annotations) to provide more accurate and semantically meaningful search results. By combining these data sources, it is possible to achieve a deeper understanding of the content and context of images, leading to improved retrieval performance. However, multimodal retrieval faces several challenges [1][2]. The conventional surveys of image retrieval cover various modalities, like text, image, speech and video. Kaur [3] conducted the surveys of image retrieval and found the deep study about their potential applications. In this type of retrieval, a system needs to be trained on multimodal data to achieve feature fusion across different modalities, for which deep learning models are employed [4]. Direct modeling and indirect modeling are the two components of multimodal methods that the researchers lay out in [5]. In an effort to categorize the multimodal retrieval approaches, the researchers in [6] attempted to do so using two types of representations: real-valued and binary. Though, recent years have seen a surge of innovative and high-tech revolutions. In the deep learning era, researchers may move away from traditional machine learning techniques. The Hybrid Cross-Modal Similarity Learning model (HCMSL), originally proposed by Zhang et al. [7], is a novel approach to

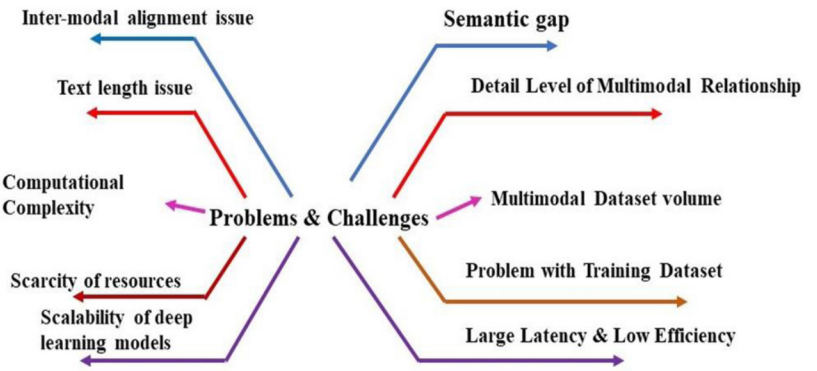

Figure 1

Multimodal Image Retrieval Challenges for Social Applications

cross-modal retrieval. Both labelled and unlabeled cross-modal pairs, as well as intra-modal pairs with the same categorization label, are intended to be adequately represented by this method's semantic information capture. As a summary of different cross-modal retrieval methods, cross-modal retrieval [8] has brought public space learning and similarity measuring. Methods based on subspaces, deep learning, hash transforms, and themes are some ways that these techniques are categorized by researchers [9].

A major problem is that there hasn't been enough study on how inter-modal semantic structures relate to intra-modal local data [10]. In their article published in Geoscience and distant Sensing, Yuan et al. [11] introduce a new approach to distant sensing text-image retrieval that crosses modalities. By employing a dual-branch network, the method captures both the overall context and fine-grained details of images and their corresponding textual descriptions [12]. Experimental results demonstrate significant improvements in retrieval performance, showcasing the potential of this technique in various remote sensing applications [13]. Figure 1 highlights a number of critical points that must be resolved. The sections that follow offer thorough introductions. This paper offers an in-depth survey and comparative analysis of various multimodal representations across different pre-training stages, providing a thorough examination of their effectiveness.

The rest of the paper as follows: section 2 deals with the gap of traditional CBIR to multimodal approaches, section 3 gives the study of multimodal techniques, section 4 deals with deep leaning modals for

effective retrieval, section 5 with multimodal datasets and experimental study and the article will be concluded by section 6.

2. Semantic gap between Traditional CBIR to MMIR

The transition from Content-Based Image Retrieval (CBIR) to Crossmodal or multimodal Image Retrieval (MMIR) marks a significant evolution in image retrieval technologies, addressing both the limitations and advancements within the field. Early CBIR systems were limited by the semantic gap, which occurred when low-level visual features did not adequately reflect the high-level semantics that people perceive, when they attempted to retrieve images based on visual features like color, texture, and shape [14]-[18]. Early CBIR approaches utilized methods like histograms and edge detection but these techniques struggled to capture complex image content accurately. Li et al. [19] has discussed the recent advancements by incorporated deep learning methods, particularly Convolutional Neural Networks (CNNs), enhancing the quality of features and bridging some of the semantic gaps Despite these improvements [20], CBIR still faces challenges related to the limited correlation between visual features and semantic understanding [21]. MMIR aims to overcome these limitations by enabling retrieval across different modalities, such as text and images. Techniques in CMIR include embedding learning, for better alignment and the use of multimodal networks that simultaneously learn from different data types [22]. These methods represent a significant advancement over traditional CBIR, offering more robust retrieval capabilities by integrating diverse data modalities. Xie et al. [23] proposed an adversarial hashing method for multimodal image retrieval. Adversarial hashing is utilized to learn hash codes that preserve the consistency of multi-modal data. Zhang et al. [24] presents a method for improving image retrieval by jointly learning attribute manipulation and modality alignment for text-image queries, enhancing retrieval accuracy across modalities. Lee et al. [25], introduced "Cosmo," a method for image retrieval that uses content-style modulation with text feedback. This approach refines image retrieval by integrating textual input to adjust image features, improving relevance and accuracy in search results. Goenka et al. [26] presented "FashionVLP," a vision-language transformer designed for fashion retrieval. Concurrently, innovations in MMIR are expected to leverage deep learning advancements, focusing on improved multimodal fusion, alignment, scalability, and real-time retrieval capabilities [27].

3. Study of Multimodal Techniques

Sentiment extraction is one of the significant advantages of multimodal image retrieval, enabling the assessment of people's emotions and the positivity or negativity polarity about health, e-commerce products, etc., aiding socially recommended applications research derived from the retrieved images. Multimodal image retrieval traces both text and image information. A novel visual and textual sentiment analysis (VITESA) [28] model defines polarity classification for multimodal data. SentiWordNet [29] classifies multimodal polarity for emotion or sentiment analysis results. Another multimodal technique, cross-modality consistent regression (CCR) [30], focuses on visual and textual sentiment analysis. CCR employs a convolutional neural network (CNN) for image-based sentiment analysis and trains a paragraph vector model for textual sentiment analysis [31]. CCR has been experimented with for multimodal sentiment analysis using Getty images and tweet images, deriving people's polarities from both image and text information. The CCR multimodal poses two different models of visual images and bag-of-words of text documents for describing the sentiment analysis. The first model uses the fine-tuned CNN for deriving the visual cues (or features) of images, and the second model generates the embedded vector and paragraph vector model for representing the text documents. Make a training of paragraph vector model by the description of image or titles of the image and learn the text feature for sentiment analysis (SA). For the multimodal data, the bi-directional multi-level attention (BDMLA) [32] model efficiently

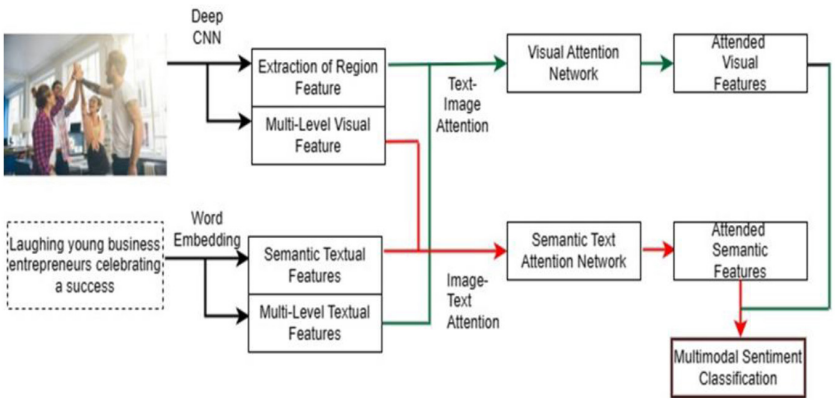

Figure 2

Architectural Diagram of BDMLA

handles the MSA with the collaborative sentiment analysis of both visual textual traits. It exploits the semantic information for two modalities, image and text, for presenting the joint visual text modelling-based sentiment classification. It focused more on emotional regions in the images and related terms in image titles or image-text pairs. Two types of networks are majorly developed in BDMLA: visual attention network and semantic attention network. Figure 2 shows the architectural diagram of BDMLA.

4. Deep Models for Effective Multimodal Image Retrieval

The deep multi-view attentive network was recently proposed in [33]; it enables MMIR with high precision using multimodal data. It uses the VGG19-imagine to extract features from the images' attractive regions or scenes. Attentive, interactive learning is essential in deep learning-based multimodal sentiment classification. The deep model concept is used to extract regions and scene features. It is developed by two modules, namely, a convolutional block attention module (CBAM) and another is a textual-guided attention module (TGAM). The CBAM aims to extract the visual cues or features from the selected regions. The aim of TGAM is to explore the emotional visual cues that are strongly associated with the textual information. In this deep multimodal framework, the cross-model fusion system is implemented by the composed features of image regions and textual sentences available with the visual image. In TGAM, text embedding is the critical component that can be created with the BERT (bi-directional encoder representation from transformers). The BERT is effectively used in many text embedding applications. It can give the best vectorization model for extracting key features of the text or sentences in terms of context, semantics, topic features, etc. Both CNN and LSTM are combined to extract text features at various embedding levels, namely phrase level, term or word level, and document level embedding. Finally, the joint features are extracted by combining the features from these three levels. Extract the visual features from the regions and scenes.

4. Deep Models for Effective Multimodal Image Retrieval

Existing multimodal techniques are executed in the Google Colab platform with a GPU and RAM of 16GB. Deep models are trained with 20 epochs only due to limited free resources in Google Colab. Deep multimodal sentiment analysis experiments are carried out using Getty

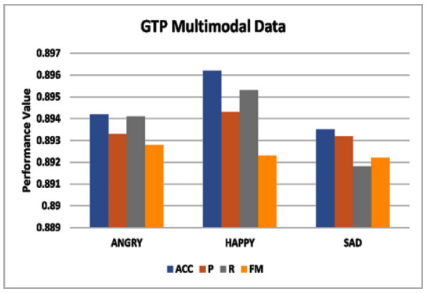

Figure 3

Performance Analysis of Deep Multimodal Sentiment Analysis for GTP Data

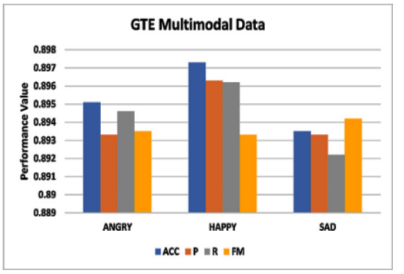

Figure 4

Performance Analysis of Deep Multimodal Sentiment Analysis for GTE Data

images with two polarities (GTP) data, Getty images with emotions (GTE) data, and multimodal tweet datasets. Multimodal sentiment analysis (MSA) experimental results in Fig. 3 for Getty images with two polarities data (negative and positive polarities) and Fig. 4 shows the experimental results for big tweets multimodal data and Getty images with three emotions data (angry, happy, and sad). The experimental results are evaluated using the accuracy (ACC), precision (P), recall (R), and f-measure (FM) [34]

From these results, it noted that the best multimodal sentiment classification results by the deep learning techniques.

5. Conclusion

Deep learning and transformer models have evidently advanced the multimodal query-based image retrieval, offered sophisticated solutions and drove significant progress in the field. In this review article, we tried to provide a thorough summary and key studies. Our coverage included feature engineering, feature fusion, model optimization, evaluation metrics for clarity. Addressing these issues is crucial for advancing the field and enhancing the precision of cross-modal retrieval systems. Despite considerable advancements, achieving optimal results in cross-modal retrieval remains challenging. Major hurdles include improving feature representation, handling complex semantic processing, aligning vision and language, developing unified architectures, optimizing models, maintaining effective performance evaluation metrics, and creating more comprehensive datasets.

References

[1] A. Vaswani, "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, p. I (2017).

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, no. 2, pp. 423-443 (2018).

[3] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review,* vol. 39, p. 100336 (2021).

[4] Y. Huo, Q. Qin, J. Dai, L. Wang, W. Zhang, L. Huang, and C. Wang, Deep semantic-aware proxy hashing for multi-label cross-modal retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 576-589 (2023).

[5] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," Proceedings of the 22nd ACM International Conference on Multimedia (2014).

[6] K. Zhou, F. H. Hassan, and G. K. Hoon, "The state of the art for cross-modal retrieval: A survey," *IEEE Access* (2023).

[7] C. Zhang, J. Song, X. Zhu, L. Zhu, and S. Zhang, "HCMSL: Hybrid cross-modal similarity learning for cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), vol. 17, no. 1s, pp. 1-22 (2021).

[8] L. I. U. Ying, G. U. O. Yingying, J. F. A. N. G., J. F. A. N., Y. H. A. O., and J. L. I. U., "Survey of research on deep learning image-text cross-modal retrieval," *Journal of Frontiers of Computer Science & Technology*, vol. 16, no. 3 (2022).

[9] J. Chen, L. Zhang, C. Bai, and K. Kpalma, "Review of recent deep learning-based methods for image-text retrieval," Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 167-172, Aug. (2020).

[10] Z. Li, H. Lu, H. Fu, and G. Gu, "Image-text bidirectional learning network based cross-modal retrieval," *Neurocomputing*, vol. 483, pp. 148-159 (2022).

[11] Z. Yuan, W. Zhang, C. Tian, X. Rong, Z. Zhang, H. Wang, and X. Sun, "Remote sensing cross-modal text-image retrieval based on global and local information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16 (2022).

[12] X. Tang, Y. Wang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Interacting-enhancing feature transformer for cross-modal remote-sensing image and text retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1-15 (2023).

[13] Q. Cheng, Y. Zhou, P. Fu, Y. Xu, and L. Zhang, "A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4284-4297 (2021).

[14] G. Sucharitha, N. Arora, and S. C. Sharma, "Medical image retrieval using a novel local relative directional edge pattern and Zernike moments," *Multimedia Tools and Applications*, vol. 82, no. 20, pp. 31737-31757 (2023).

[15] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," arXiv preprint arXiv:1706.06064 (2017).

[16] G. Sucharitha and R. K. Senapati, "Local quantized edge binary patterns for colour texture image retrieval," *Journal of Theoretical & Applied Information Technology*, vol. 96, no. 2 (2018).

[17] I. M. Hameed, S. H. Abdulhussain, and B. M. Mahmmod, "Content-based image retrieval: A review of recent trends," *Cogent Engineering*, vol. 8, no. 1, p. 1927469 (2021).

[18] X. Li, J. Yang, and J. Ma, "Recent developments of content-based image retrieval (CBIR)," *Neurocomputing*, vol. 452, pp. 675-689 (2021).

[19] M. Alrahhal and K. P. Supreethi, "Content-Based Image Retrieval using Local Patterns and Supervised Machine Learning Techniques," 2019 Amity International Conference on Artificial Intelligence (AIC-

AI), Dubai, United Arab Emirates, pp. 118-124 (2019), doi: 10.1109/AICAI.2019.8701255.

[20] S. Qian, Y. Guo, J. Liu, X. Zhang, and Q. Wu, "Adaptive label-aware graph convolutional networks for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 24, pp. 3520-3532 (2021).

[21] N. Arora, G. Sucharitha, and S. C. Sharma, "MVM-LBP: Mean–Variance–Median based LBP for face recognition," *International Journal of Information Technology* (2023).

[22] M. Hosseinzadeh and Y. Wang, "Composed query image retrieval using locally bounded features," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020).

[23] D. Xie, X. Zhang, Y. Li, J. Chen, L. Wang, and W. Li, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 29, pp. 3626-3637 (2020).

[24] F. Zhang, M. Xu, Q. Mao, and C. Xu, "Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval," Proceedings of the 28th ACM International Conference on Multimedia, pp. 3367-3376 (2020).

[25] S. Lee, D. Kim, and B. Han, "Cosmo: Content-style modulation for image retrieval with text feedback," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).

[26] S. Goenka, Z. Zheng, A. Jaiswal, R. Chada, Y. Wu, V. Hedau, and P. Natarajan, "FashionVLP: Vision language transformer for fashion retrieval with feedback," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14105-14115 (2022).

[27] N. Perveen, D. Roy, and C. K. Mohan, "Spontaneous expression recognition using universal attribute model," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5575-5584 (2018).

[28] M. Paolanti, C. Kaiser, R. Schallner, E. Frontoni, and P. Zingaretti, "Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks," in Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania,

Italy, Sep. 11-15, Part I, Springer International Publishing, pp. 402-413 (2017).

[29] N. Medagoda, S. Shanmuganathan, and J. Whalley, "Sentiment lexicon construction using SentiWordNet 3.0," in 2015 11th International Conference on Natural Computation (ICNC), *IEEE* (2015).

[30] Q. You, J. Luo, H. Jin, and J. Yang, "Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia," in Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pp. 13-22 (2016).

[31] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 879-887 (2019). doi: 10.1109/TMM.2018.2887021.

[32] J. Xu, F. Huang, X. Zhang, S. Wang, C. Li, Z. Li, and Y. He, "Visual-textual sentiment classification with bi-directional multi-level attention networks," *Knowledge-Based Systems*, vol. 178, pp. 61-73 (2019).

[33] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, "Hi-net: Hybrid-fusion network for multi-modal MR image synthesis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2772-2781 (2020).

[34] E. P. Ijjina and C. K. Mohan, "Human action recognition in RGB-D using motion sequence and deep learning," *Pattern Recognition*, vol. 72, pp. 504-516 (2017). doi: 10.1016/j.patcog.2017.07.01.

Received November, 2024