Computer Science Review 39 (2021) 100336

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Review article

Comparative analysis on cross-modal information retrieval: A review

Parminder Kaur [a,*], Husanbir Singh Pannu [a], Avleen Kaur Malhi [b]

[a] Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, India
[b] Department of Computer Science, Aalto University, Finland

ARTICLE INFO

Article history:
Received 20 September 2020
Received in revised form 26 November 2020
Accepted 27 November 2020
Available online 8 December 2020

Keywords:
Cross-modal
Multimedia
Information retrieval
Data fusion
Comparative analysis

ABSTRACT

Human beings experience life through a spectrum of modes such as vision, taste, hearing, smell, and touch. These multiple modes are integrated for information processing in our brain using a complex network of neuron connections. Likewise for artificial intelligence to mimic the human way of learning and evolve into the next generation, it should elucidate multi-modal information fusion efficiently. Modality is a channel that conveys information about an object or an event such as image, text, video, and audio. A research problem is said to be multi-modal when it incorporates information from more than a single modality. Multi-modal systems involve one mode of data to be inquired for any (same or varying) modality outcome whereas cross-modal system strictly retrieves the information from a dissimilar modality. As the input–output queries belong to diverse modal families, their coherent comparison is still an open challenge with their primitive forms and subjective definition of content similarity. Numerous techniques have been proposed by researchers to handle this issue and to reduce the semantic gap of information retrieval among different modalities. This paper focuses on a comparative analysis of various research works in the field of cross-modal information retrieval. Comparative analysis of several cross-modal representations and the results of the state-of-the-art methods when applied on benchmark datasets have also been discussed. In the end, open issues are presented to enable the researchers to a better understanding of the present scenario and to identify future research directions.

Contents

* Corresponding author.
E-mail addresses: pkaur60_phd18@thapar.edu (P. Kaur), hspannu@thapar.edu (H.S. Pannu), avleen.malhi@aalto.fi (A.K. Malhi).

https://doi.org/10.1016/j.cosrev.2020.100336
1574-0137/

1. Introduction

When we fail to understand the contents of an image embedded in a text, figure captions, and referral text often help. Just by looking at a figure, a person might not be able to understand it exactly but with the help of collateral text, it can be understood efficiently. For instance, when we see a volleyball picture (Fig. 1), we may not be able to understand or know about the volleyball game. However, the picture can be completely understood with the help of collateral text (such as caption, figure reference, and related citation) describing the volleyball game. This implies that information from more than one source is beneficial in further understanding of things and also helpful in better information retrieval. This is where cross-modal data fusion and retrieval come into the picture.

Recently, cross-modal retrieval has gained a lot of attention due to the rapid increase in multi-modal data such as images, text, video, and audio. The term modality represents a specific form in which data exists and it is also associated with a sensory perception such as vision and hear modalities which are major sources of communication and responsiveness in human beings and animals. The data consisting of more than one modality is known as multi-modal data. It has the characteristic of high-level semantic homogeneity and low-level expressive heterogeneity such as the same thing having diverse representations. Different forms of representation help people better understand things as illustrated in the volleyball example above. While searching for something, people often want to get accurate results in different forms which create a need for an efficient multi-media information retrieval platform. Classic approaches to information retrieval are of uni-modal nature. Uni-modal means information derived just from one channel, such as only from images or only from the text (but not both). For example, only the text query is used for information search and retrieval from a text repository. This retrieval approach is of the least use these days when enormous multimedia data is being generated. Cross-modal and multi-modal systems, on the other hand, are able to link more than one modalities such as image, text, audio, and video. In cross-modal, input query mode and resultant mode are dissimilar. For example, query text for related images and query image for related text. However, the resultant mode can be similar to the query mode in a multi-modal system. For example, query text to retrieve related images and matched text. Cross-modal and multi-modal are explained using a simple example in Fig. 2 where + represents both text and images can be retrieved using an image query and vice versa in multi-modal approach.

Therefore, the fundamental idea of cross-modal is to integrate numerous modes of information to derive better results than just one channel. For instance, an image–text cross-modal system integrates textual information along with an image which is known as *image annotation*. Vice-versa, it also queries text keywords to retrieve images, known as *image retrieval*. In simple words, image annotation is a process of explaining an image with appropriate linguistic cues. It is useful in knowledge transfer sessions for application areas such as medical science, military, business, education, and sports to name a few. For example, a CT scan is known to the radiologist but not to an intern or a patient. Therefore, the expert has to explain it using proper terminology by pointing out key areas on the given image. Image retrieval is a process of retrieving an appropriate image from the database as per the user query, for instance, with text keywords. With the evolution of the semantic web and huge data repositories, a major challenge comes into the picture which is effective indexation and retrieval of both still and moving images and the identification of key areas inside the images. An image cannot be expressed completely just by using visual features only as they under-constrain the information contained in it. Visual features of an image include color distribution, texture, shape, and edges. Typically, image retrieval systems make use of images and the corresponding text/keywords for indexing and retrieving images using both keywords and visual features of the image. Cross-modal image retrieval aims to use text for retrieving relevant images related to the text.

1.1. Motivation

Cross-modal learning has become tremendously popular because of its effective information retrieval capability. Numerous cross-modal representation and retrieval methods have been proposed by researchers to resolve the issue of cross-modal retrieval considering several modalities. Various appealing surveys have been introduced which summarizes the work done in this field. Image and text are the highly utilized modalities and a number of articles on cross-modal retrieval have been published considering these. However, there is no proper survey mainly focusing on the image–text cross-modal retrieval techniques. The objective of this article is to conduct a comprehensive review of cross-modal retrieval which incorporates image and text modalities, the main concerns of which are different from previous surveys and reviews. So, the motivation behind this review article is:

1. Lack of a full-fledged review article on image and text modalities.
2. To present various challenges and open issues in the cross-modal retrieval field.
3. Image and text modalities are the basic and highly utilized modalities, however, we are still away from achieving an ideal level in their cross-modal retrieval process.

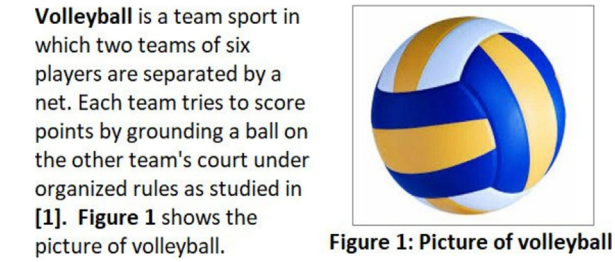

Fig. 1. An example of a volleyball image and collateral text in the form of the caption, figure reference, and related citation.

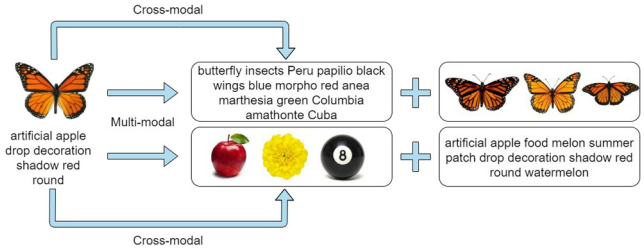

Fig. 2. An illustration of information retrieval in cross-modal and multi-modal system.

1.2. Related surveys

Existing literature reviews related to cross-modal information retrieval have presented the topic quite well to the research community. [1] presented an overview of cross-modal retrieval in 2016, however, it does not comprise several significant works proposed in recent years. In [2], authors have presented numerous multi-modal techniques, but their focus is only on techniques based on machine learning. [3] is a contemporary survey, however, it presents a brief study of cross-media retrieval methods compared to the vastness of the topic. An overview of different cross-media retrieval techniques incorporating miscellaneous modalities has been provided in [4]. [5] article only explore various cross-media retrievals with joint graph regularization. The focus of [6] is on cross-media analysis and reasoning and the various analysis methods rather than cross-media retrieval. [7] has provided a survey on cross-media image and text information fusion where the main focus is on analyzing two methods of image and text associations.

Table 1 shows the comparison of the current survey with the existing reviews related to cross-modal learning sorted year wise. Comparison is performed on the basis of the domain, different modalities incorporated in the paper, comparative analysis, challenges, open issues, benchmark datasets, and evaluation metrics. It can be seen in the table that only one survey is focusing on image and text modalities but their main concern is an image–text association and not cross-modal retrieval. A blank cell in the table implies that the information is missing for that particular column and ✓means that it is present in the article. *Domain* column specifies the main focus of the article and *all* value under *Mode* column means that the article is not particularly focusing on any two or three modalities rather it is talking about the whole multi-media. *Comparative analysis* depicts whether the comparison among techniques has been performed quantitatively or qualitatively.

1.3. Contributions

The significant contributions of this paper are as follows:

1. This review focuses to present a summary of recent progress in cross-modal retrieval considering text and image (image-to-text and text-to-image). It comprises several novel works and references which are absent in previous surveys. It will act as a valuable resource for beginners to get acquainted with the topic.
2. A broad classification of various cross-modal approaches has been presented and difference among them is also discussed.
3. It provides information regarding various prominent benchmark datasets and evaluation metrics utilized for cross-modal method performance estimation.
4. It presents a comparative analysis of diverse cross-modal representation techniques when applied on benchmark datasets. This analysis will be highly useful for future research.
5. The article summarizes various challenges in the field of cross-modal retrieval and open issues to work upon by future researchers.

1.4. Article organization

This article starts with an introduction of cross-modal retrieval in *Section* 1 which includes motivation for the survey, contributions, comparison with existing surveys, article road map, and organization. An appropriate review methodology (*Section* 2) has been shadowed in writing the proposed survey which incorporates five subtopics: research questions, sources of information, search criteria, data extraction, and publication metrics. The inception of cross-modal retrieval, its general architecture, applications, observed challenges in the process, and the initial related articles are presented in concert under the background section (*Section* 3). *Section* 4 discusses about the diverse cross-modal representation and retrieval techniques which are broadly classified into real-valued and binary techniques. The literature related to these techniques has also been included in this section. The famous image–text benchmark datasets which have been widely used by the researchers in the cross-modal field have been presented in *Section* 5. *Section* 6 is of comparative analysis which introduces different performance evaluation metrics along with a comparison of various cross-modal retrieval methods. A summary of several state-of-the-art cross-modal retrieval works has been demonstrated with the use of tables in *Section* 7. The miscellaneous open issues in cross-modal retrieval domain have been discussed in *Section* 8. Finally, *Section* 9 culminates the survey with the conclusion. Fig. 3 depicts the road map of the article.

2. Review methodology

The categorical survey technique described in this research article has been obtained from the technique described by Kitchenham et al. [8,9]. Distinct stages used in this review are: to create a review technique, planning an exhaustive survey, executing the survey, comparison of results, comparative result analysis, and exploring open issues. The review technique employed in this categorical survey is described in Fig. 4.

2.1. Research questions

A number of vital areas required to be considered in the case of cross-modal retrieval are summarized in the following research questions:

1. *RQ1*: What is the need of cross-modal retrieval?
   AS1: The results achieved in information retrieval can be highly improved when information from more than one mode is incorporated into the process. [Details in *Section* 1]

Table 1
Comparison of the proposed survey with existing surveys.

Sr. No.	Article	Year	Domain	Mode	Comparative analysis	Chal-lenges	Open issues	Datasets	Eval. metrics
1	Priyanka [7]	2013	Cross-media text and image information fusion	Image and text					
2	Wang et al. [1]	2016	Cross-modal retrieval	all	Quantitative	√	√	√	√
3	Peng et al. [4]	2017	Cross-media retrieval: concepts, benchmarks, methodologies and challenges	all	Quantitative	√	√	√	√
4	Peng et al. [6]	2017	Cross-media analysis and reasoning advances and directions	all		√	√		
5	Baltruvsaitis et al.[2]	2018	Multimodal machine learning	all	Qualitative	√			
6	Monelli and Bondu et al.[5]	2018	Joint graph regularization based semantic analysis for cross-media retrieval	all	Qualitative				
7	Monelli and Bondu [3]	2019	Cross-media feature retrieval and optimization	all		√	√		
8	**Proposed survey**	–	Cross-modal retrieval considering image and text modalities	Image and text	Qualitative and quantitative	√	√	√	√

Table 2
List of journals represented by **X** in Fig. 6, having a publication count of 1.

Publisher	Journals	
Elsevier	Alexandria Engineering; Computer Vision and Image Understanding; Digital Signal Processing	
	Information Sciences; Procedia Computer Science; Signal Processing: Image Communication	
Hindawi	Mathematical Problems in Engineering; Mobile Information Systems	
IEEE	Journal of Selected Topics in	Applied Earth Observations and Remote Sensing; Signal Processing
	IEEE Signal Processing Magazine	
	Transactions on	Automation Science and Engineering; Biometrics, Behavior, and Identity Science; Circuits and Systems for Video Technology
		Dependable and Secure Computing; Geoscience and Remote Sensing; Industrial Electronics
		Industrial Informatics; Neural Networks and Learning Systems
IET	CAAI Transactions on Intelligence Technology	
	Image Processing	
Springer	Computer Vision; Computer Science and Technology	
	Neural Computing and Applications; Pattern Analysis and Applications	
	Signal, Image and Video Processing; Soft Computing	
Taylor & Francis	Intelligent Automation & Soft Computing; The Imaging Science	
Other	International Journal of Signal Processing, Image Processing and Pattern Recognition	

2. *RQ2*: What is the background of the word cross-modal?
   AS2: Cross-modal learning is inspired from working of human brain. [Details in *Section* 3]
3. *RQ3*: What are the different challenges faced during the process of cross-modal retrieval?
   AS3: Handling of huge multi-modal datasets, heterogeneous modalities and others. [Details in *Section* 3.3]

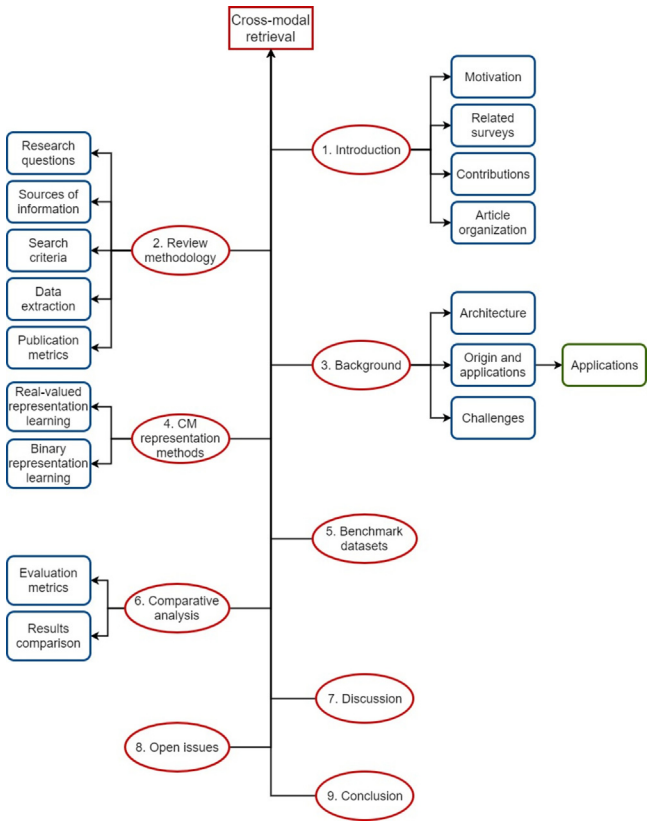

Fig. 3. Road map for the article.

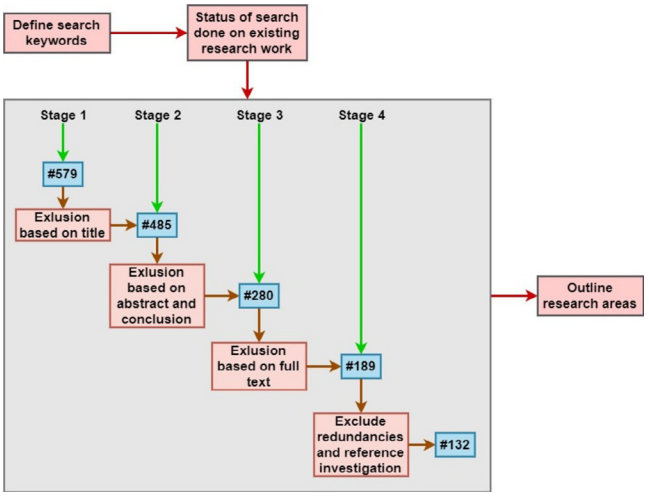

Fig. 4. Shortlisting of the research articles starting from title, abstract/conclusion, full text and redundancy.

4. *RQ4*: What are various applications of cross-modal retrieval?
   AS4: Emotion recognition system, biomedical image retrieval, and spoken to sign language transcription to name a few. [Details in *Section* 3.2.1]
5. *RQ5*: Which state-of-the-art methods have been proposed recently for multi-modal data representation and retrieval and which image and text descriptors have been used in combination?
   AS5: The popular multi-modal data representation method is CCA, SIFT and LDA are famous methods utilized for image and text representation respectively. [Details in *Section* 4 and *Table* 18]
6. *RQ6*: What are the various existing prominent image–text benchmark datasets?
   AS6: NUS-WIDE, Wikipedia and MIRFlickr 25k are few popular datasets. [Details in *Section* 5 and *Table* 7]
7. *RQ7*: What are the primary evaluation metrics utilized by researchers and comparison of miscellaneous techniques on different benchmark datasets?
   AS7: MAP and PR curve are commonly accepted evaluation metrics by research community. [Details in *Section* 6]
8. *RQ8*: What are the open issues in the field of cross-modal retrieval?
   AS8: Lack of huge multi-modal datasets, restricted annotations and diversity requirement in dataset composition are few open issues. [Details in *Section* 8 and *Table* 17]

Google Scholar	www.scholar.google.com
IEEE eXplore	www.ieeexplore.ieee.org
Springer	www.springer.com
ACM digital library	www.acm.org/dl
Science direct	www.sciencedirect.com

Fig. 5. The electronic databases used in the survey.

2.2. Sources of information

We searched broadly in electronic database sources as recommended by Kitchenham et al. [8,9]; the electronic databases used for searching are given in Fig. 5.

2.3. Search criteria

The survey conducted contains the literature review of the qualitative and quantitative research articles from the year 2010 to 2020 in the English language. In this article, we have included research papers from peer-review journals, symposiums, conferences, technical reports, and workshops. The exclusion criteria used in the search is given in Fig. 4. An individual search was applied to a few articles from Springer, Elsevier, and IEEE to name a few to cross-check the electronic searching. There were 579 articles gathered from the search which were then further reduced to 485 based on the titles of the articles. The exclusion was done based on the titles as the titles which were relevant to image–text cross-modal were kept in the list and titles which seemed out of the scope of the area were excluded (such as papers based on other cross-modal retrievals were excluded). The number was further reduced to 280 based on the abstract and conclusion of the article. The abstracts/conclusions of all the papers were read and relevant papers to cross-modal retrieval considering image and text were selected among all other papers. Finally, 189 articles were selected based on their full text as the papers whose technique was not relevant to our domain were excluded. Then, these 189 articles were examined thoroughly to give the final list of 132 research papers with the help of reference investigation and redundancies to eliminate common challenges based on inclusion and exclusion criteria.

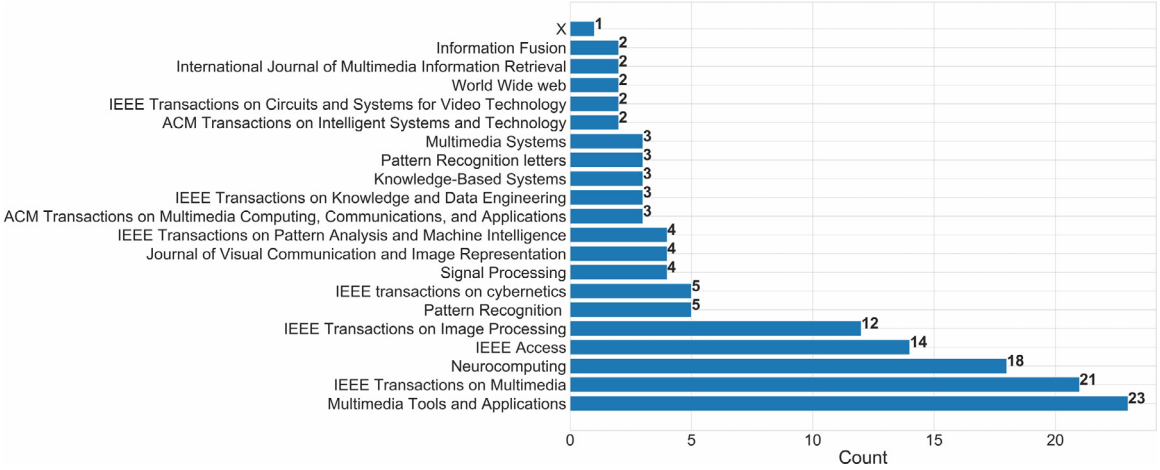

Fig. 6. Journal publications count in the area of Cross-modal retrieval during last decade. Journals represented by **X** are given in Table 2.

2.4. Data extraction

Many problems were faced in extracting relevant information from the sources specified. Various authors were contacted for finding the in-depth knowledge of research if required. Our review used the following procedure for data extraction:

- The data from 132 research articles was extracted by one of the authors after a detailed review.
- Another author then cross-checked the review results extracted.
- If any conflict arose during cross-checking, then a compromise meeting was held by authors to resolve the conflict.

The aim of this review is to find the available research in image–text cross-modal retrieval. Most of the research articles on cross-modal retrieval are published in a comprehensive variety of referred journals and conference proceedings.

2.5. Publication metrics

This section involves the publication metrics related to cross-modal retrieval. Fig. 7 presents a chart of year-wise publications from year 2010 to 2020 which has been obtained by executing the search string *allintitle: retrieval "cross modal" OR "cross media" OR "multi modal" OR "multi media"* on Google Scholar as per year. It can be inferred from the chart that the publication count in the field has increased overall.

Fig. 6 displays the publication count of prominent journals in the field of cross-modal retrieval during last decade. Journal *Multimedia Tools and Applications* has the highest publication count of 23. Here, **X** represents all journals having publication count of 1. They are given in Table 2. Fig. 8 demonstrates the cross-modal retrieval journal publication count geographically on the world map during last decade. *China* has the maximum publication count with a score of 174. *India* and *US* are on second and third position with a score of 15 and 11 respectively.

3. Background

The inception of the terms *cross-modal* and *multi-modal* is in neurology and are inspired from multi-sensory integration inside brain [10,11]. We often need to understand images of objects/scenes through the use of phrases because image does not contain all the relevant information. Thus, we use one modality of communication to compensate for the absence of information in another mode [12] which implies co-relating text and image.

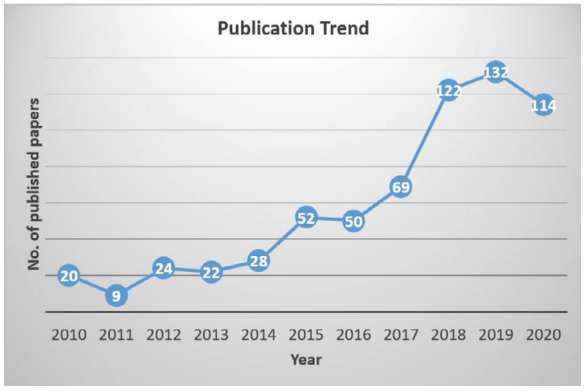

Fig. 7. Cross-modal retrieval article publishing trend.

Fig. 8. Geographical representation of country wise journal publication count in last decade.

In simple terms, cross-modal or multi-modal is linking of more than one modality such as text, image and video. The process of using one modality to retrieve related information in other modality is known as cross-modal retrieval. Retrieval of text using an image is called image annotation and retrieval of an image using text is known as image retrieval.

Fig. 9 shows a lucid comparison of various modalities utilized inside the brain and computer. For instance, sight modality for the brain is collated with image modality for computer, hear is collated with audio, and so on. Humans get familiar with the outside world using multiple sensory channels where each channel provides a distinctive impression of the environment [13]. Each

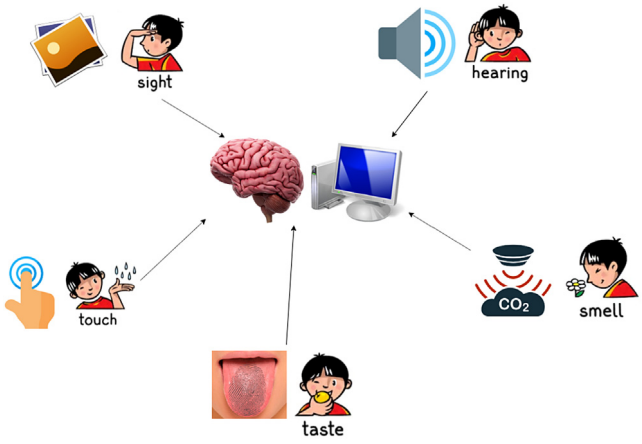

Fig. 9. Comparison of computer and brain, based on different modalities which are integrated inside them to make a decision. For instance, image modality in a computer is similar to vision modality in the brain, tongue biometric is similar to taste and so on.

sense or sensory modality seems to exercise separately to interact with the environment and acquire knowledge, however, the information received from all the sensory channels is integrated by the brain into an extensive awareness regarding the outside world [14]. As an example, when it is difficult to understand a person's articulation then the interlocutor will automatically start observing other modes of expression such as mouth movements and facial/body expressions. In a similar fashion, only one modality is not enough to understand an incident/object for decision making. So, computers can be evolved to the next level by integrating information from diverse modalities to achieve better results than just one information channel. Current generation computers are not capable to perform cross-modal computation efficiently due to their existing structure (separation of memory and processor unlike brain neurons [15]) and complexity involved in the cross-modal learning phenomenon. Hence, cross-modal learning is an attempt to make the computer evolve to the next level.

3.1. Architecture

Fig. 10 demonstrates the general framework of a cross-modal retrieval system. Four modalities such as text, image, video, and audio are shown in the figure as an example. Typically, the initial form of data contains noise which affects the overall building and accuracy of the system. Pre-processing is performed on the data to remove that noise and to make it appropriate for further processing on it. The second step is to represent each modality separately by doing a feature extraction process using varied algorithms such as BoW, SIFT, and CNN depending upon the multi-modal data. As per the multi-modal representations, common representations for diverse modalities are learned using correlation modeling. In the end, this common representation enables the cross-modal retrieval process by appropriate ways of data indexing, summarization, and ranking.

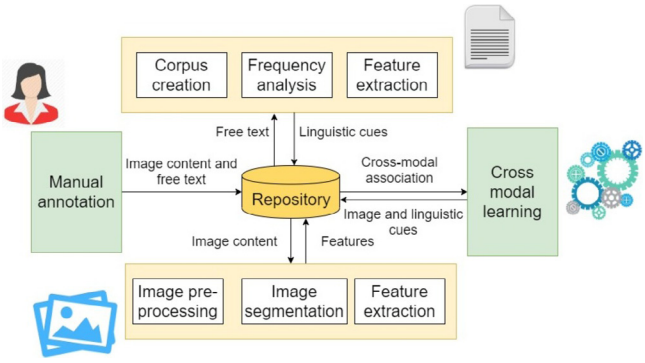

Fig. 11. General topological architecture of image–text cross-modal system.

Fig. 11 shows the general topological architecture of cross-modal image annotation and retrieval system. It defines the relative locations of each entity in the rectangular boxes and process flow through directed arrows. Working of each entity is briefed as follows:

1. *Manual annotation*: Manual annotation implies the labeling of unlabeled images or providing an appropriate explanation for each image by an expert.
2. *Repository*: A repository refers to a central location for the storage and management of data consisting of images and text. It is the common location for fetching and saving data for all the entities present in the system. The prepared and organized data after an annotation is put into the repository for further analysis. The image and textual data are picked up by image and text analysis modules separately for analysis and then the extracted features are put back in the repository. The cross-modal learning module

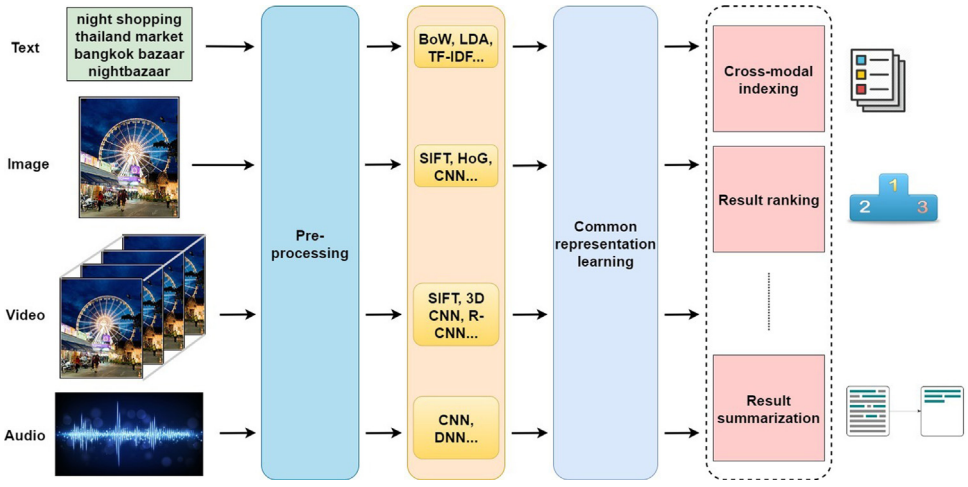

Fig. 10. General framework of cross-modal retrieval process.

is connected to the repository so that when a query is fed into the module, it can retrieve the related result from the repository module.

3. *Image analysis*: This module consists of three sub-modules which are as follows:
   - Image pre-processing: It involves the cleaning of noisy images, improving the quality of blur images, and image resizing.
   - Image segmentation: It is a process of segregation of an image into several small segments such as a group of pixels which makes the image representation easier to analyze.
   - Feature extraction: It consists of withdrawing the useful features from an image which uniquely identifies it.
4. *Text analysis*: This is further divided into three steps:
   - Corpus creation: It includes text pre-processing, typically consisting of noise and stop words removal. After pre-processing, the final word corpus is created.
   - Frequency analysis: It involves assigning a frequency to the words in the corpus.
   - Feature extraction: Like image feature extraction, text feature extraction identifies the vital features from the text which differs it from other text. Few feature extraction methods include Bag-of-Words (BoW), TF–IDF, and weirdness coefficient.
5. *Cross-modal learning*: This is the most important module in the architecture and is the final system which is used for image annotation and retrieval purpose. When a text query is fed into it, it fetches the matched text and related images from the repository and returns them to the user.

3.2. Origin and applications

Most of the initial works which are inspired by cross-modal human behavior are related to the integration of acoustic and visual modalities [16–18]. Numerous works proposed in the 80s and 90s are influenced by the McGurk effect [19]. As per this effect, an illusion happens when an acoustic element of one sound is combined with the visual element of another sound which leads to the perception of a third sound. After getting motivation from previous researches on cross-language information retrieval (CLIR) and spoken document retrieval (SDR) *Thijs Westerveld* presented one of the earliest researches on cross-modal retrieval considering image and text [20,21]. In both of his works, he proposed the use of the Latent Semantic Indexing (LSI) method for cross-modal image–text retrieval. This method builds a common multi-modal semantic space to represent images and text simultaneously which benefits the retrieval of related text using an image and vice-versa.

3.2.1. Applications

Authors have introduced numerous techniques for cross-modal information retrieval considering miscellaneous modalities and applied them in varied applications. Summary of prominent and recent applications exploiting various modalities is stated in the Table 3. As this survey is focused on image and text modalities, the prominent works and representation techniques related to those are mentioned in Section 4. Few applications are described as below and also presented in Fig. 12:

1. *Face-voice matching and retrieval*: Studies reveal that human appearances are linked to their voices and humans have the tendency to recognize the association of voice and face. As an example, after hearing a voice, humans can easily identify the gender of the person and the approximate age. Inspired by this, a cross-modal framework for face voice matching and retrieval is proposed [22]. In addition to the framework, a novel face-voice dataset has been constructed from Chinese speakers and formed a data collection tool as well.

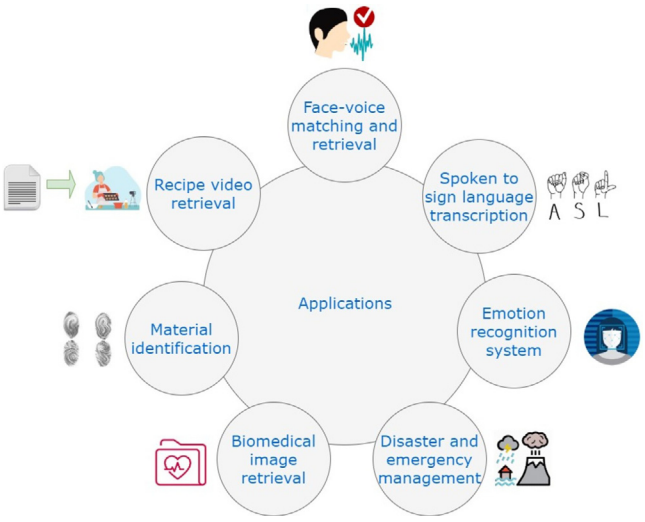

Fig. 12. Applications of cross-modal retrieval.

2. *Spoken to sign language transcription*: Cross-modal data fusion has been applied to explore the spoken language to sign language and vice versa [23]. Sign language is helpful to assist the communication with hard-of-hearing or deaf people and the proposed methodology has introduced *How2Sign* technology which is based upon the American Sign Language dataset.
3. *Emotion recognition system*: *Psycho-linguistics* is the study of the mental aspects of speech and language. As per the studies on human communication, people start adjusting their behavior by imitating the expressions, gestures, and speaking style of their interlocutor which is known as *entrainment*. Entrainment presence in cross-modal settings is investigated in [24] and its effect on multi-modal emotion recognition system. Moreover, cross-speaker dependence and cross-modality have also been inquired about. After empirical analysis, it has been found that there is a strong relationship between acoustic and facial features of one person with the emotional state of other and speakers show alike emotions specifying powerful mutual influence in their expressive behaviors 72% of the time. It has been found that there is a robust dependence among heterogeneous modalities across interlocutors and the emotional state of an interlocutor can be identified from the information provided by the other interlocutor.
4. *Disaster and emergency management system*: A large number of disaster videos, images, and news are uploaded and searched on social media regularly. These multimedia can be utilized as sensors to extract important information about the disasters. Images and text have been associated by [25] to extract prominent phrases related to floods. Bag of words model for text features and Speeded-up Robust Features (SURF) for image feature extraction has been used. For the integration of image and text, analysis has been done using a proposed novel method. Two flood event corpora were used for experiments (a) US Federal Emergency Management Agency media library, and (b) public

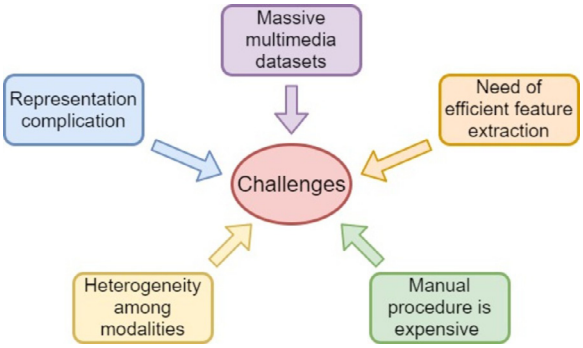

Fig. 13. Challenges in cross-modal retrieval.

Facebook groups and pages for the flood and the aid (in German).

5. *Biomedical image retrieval*: Considering the growth of the healthcare industry, important text and images keep on hiding under the inessential data which makes it hard to retrieve the relevant information. Biomedical articles often contain annotation markers or tags such as letters, stars, symbols, or arrows in their figures to spot the highlights. These markers are also correlated with the image captions and text in the article. Identification of the markers becomes important to extract the Region of Interest (ROI) from images. A novel technique has been proposed in [26] with the combination of rule-based and statistical image processing ways for localizing and annotating the medical image regions or ROIs. Moreover, a cross-modal image retrieval technique has been implemented based on ROI identification and classification.
6. *Material identification*: A combination of visual and audio can be utilized for identifying a particular material. In [27], authors are identifying a wooden material based on sound and its image. ELM-LRF method is used for feature extraction from images and audio. For cross-modal data representation, CCA, MCCA (Mean CCA), and CCCA (Cluster CCA) approaches are utilized and out of which CCCA is found to be the best among the three.
7. *Recipe video retrieval*: A video of a recipe can be retrieved from a textual recipe. A cross-modal fusion sub-network is proposed in [28] for obtaining video from a written recipe. It learns both independent and collaborative dynamics which enhances the associated representation of videos and recipes. Co-attention network utilized for explicitly emphasizing the cross-modal interactive features between recipe and video.

3.3. Challenges

The main challenges observed in the process of cross-modal retrieval are represented in Fig. 13 and defined as follows:

1. *Massive multimedia datasets*: Sophisticated content retrieval has become a challenge in ever-growing multimedia data volume over the internet [47]. Thus the model efficiency and accuracy suffer from the relevant feature extraction, selective data retention by removing redundancy while taking care of language syntax, and semantic interpretations. It is also challenging to store and retrieve data in real-time cross-modal systems to serve a useful purpose and claim the automatic semantic application.
2. *Heterogeneity among modalities*: The ever increasing size of multimedia data on social media every day creates a bottleneck for efficient information retrieval [48]. Mobile devices and social websites such as Twitter, Facebook, and Flickr are generating a variety of heterogeneous data which is semantically different and cannot be compared directly in their initial form. It is required to reduce the *semantic gap* among miscellaneous modalities so that they can be compared/matched with each other to find similarities. Semantic gap refers to the difference between low-level features and high-level concepts. The nature of data distributions, noise/artifacts, and key features involved in various modalities are subtle and prone to errors while orchestrating them for mutual information retrieval. Therefore multimedia data and massive size present a challenge.
3. *Manual procedure is expensive*: Most of the data that is found on the Internet these days is either not annotated or inaccurate. It is quite difficult to annotate the raw data (images for example) manually by an expert due to its massive volume and diversity. Hence it is needed to leverage this manual process for an automatic replacement which is comparatively accurate [49].
4. *Need of efficient feature extraction*: Choosing optimal feature extraction method for underlying multi-modal data is still an open question [50,51]. How effectively a modality has been represented through its feature vectors eventually affects the overall model quality and reliability. There is a traditional trade-off between the time complexity and model validation accuracy so the art is to find mutual equilibrium. With effective feature extraction best practices, identifying similarities between modalities will be easier and efficient.
5. *Representation complication*: In the case of multiple modalities, the basic problem is about coherent representation and synchronization among various modalities which are often not complementary and thus carry redundancy. Thus, it is important to have precise spectrum representation with maximum information gain and no redundancy. For example, in the case of image and text, images can be represented in spatial or spectral while the text is symbolic and dependent upon grammar rules and cultural norms [2].

4. Cross-modal representation and retrieval techniques

Cross-modal representation techniques can be broadly classified into two categories: (a) Real-valued representation and (b) Binary representation. In *real-valued representation learning*, the learned common representations of diverse modalities are real-valued. However, in *binary representation learning*, diverse modalities are mapped into a common hamming space. Cross-modal similarity searching is faster in binary representation, so the retrieval process also becomes faster. However, the retrieval accuracy becomes less in binary representation as the information is lost because representation is encoded to binary codes. Prominent cross-modal learning methods and related works are presented in the following sub-sections. Fig. 14 presents a taxonomy of cross-modal retrieval methods. Table 4 shows the list of acronyms used in this article. Fig. 15 presents the literature classification utilized in this survey.

4.1. Real-valued representation learning

This section presents the information regarding various real-valued representation learning methods and their application on different datasets. Fig. 16 presents the evolution of real-valued representation learning methods in recent years.

Table 3
Various application areas of cross-modal information retrieval along with modalities used, year, and reference. Modalities' representation: A-audio, I-image,3D-three dimensional, T-text, V-video, L-location.

Sr.	Modalities	Applications	Year	Ref.
1	A, I	Face-voice matching and retrieval	2019	[22]
		Material identification	2019	[27]
2	A, I, 3D	Image–audio-3D retrieval in multiple concepts	2013	[29]
3	A, I, T	Multiple distinct category image–audio and image–text retrieval	2019	[30]
4	A, I, T, V, 3D	Image–text–audio-video cross modal retrieval in various concepts	2013	[31]
5	A, T	Text–audio retrieval in audios from Freesound library	2019	[32]
		Text–audio retrieval in multiple concepts from Spotify, YouTube and Musixmatch sources	2019	[33]
6	A, V	Audio–visual retrieval in distinct categories of YouTube and Google audio set videos	2019	[34]
		Emotion recognition systems	2013	[24]
		Human behavior analysis	2018	[35]
7	A, V, L	Audio–visual-location cross-modal retrieval in distinct categories	2011	[36]
8	A, V, T	Youtube video categorization	2020	[37]
9	I, T	RoI identification and classification in CT scan images	2014	[26]
		Image retrieval in different categories	2014, 2019	[38,39]
		24 distinct category image–text retrieval	2012	[40]
		Disaster and emergency management	2016	[25]
		Image–text retrieval in various categories	2017, 2018	[41,42]
10	I, T, V	Image–text and video–text retrieval in multiple categories	2015	[43]
		Video, image and text retrieval in video lectures	2014	[44]
11	T, V	Multiple concepts' video annotation	2011	[45]
		Cooking activities' video annotation, videos' temporal activity localization evaluation, personal videos' annotation	2019	[46]
		Cooking recipe retrieval	2019	[28]

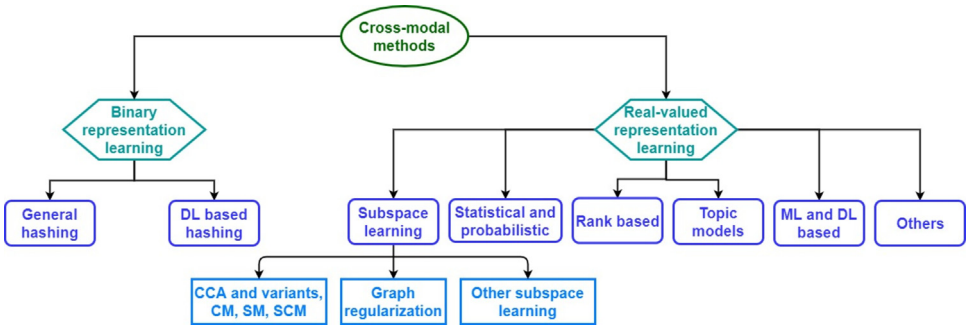

Fig. 14. Taxonomy of cross-modal retrieval methods.

4.1.1. Subspace learning

Subspace learning plays a vital role in cross-modal information retrieval. Diverse modalities have different representation features as well as they are located in diverse feature spaces [52]. The modalities can be mapped to common isomorphic subspaces from old miscellaneous spaces by learning potential common subspaces (as shown in Fig. 17).

CCA and its variants, CM, SM and SCM. CCA is the most popular unsupervised technique of subspace learning which was introduced by Hotelling [53] in 1936. The principal logic behind this technique is to find the pair of projections for diverse modalities such that the correlation between them is maximized [54]. CCA can be recognized as an issue of identifying the basis vectors for two group of variables aiming to mutually maximize the correlation between variables' projections onto the basis vectors [55]. Let $\langle \cdot, \cdot \rangle$ represents the euclidean inner product of vectors p, q which is equal to $p'q$, where A' is the transpose of a vector or matrix A. Let (p, q) denotes a multivariate random vector and its sample instances as $S = ((p_1, q_1), \ldots, (p_n, q_n))$. S_p represents $(p_1, \ldots, p_n)$ and $S_q = (q_1, \ldots, q_n)$, consider defining a new coordinate for p by choosing a direction d_p and projecting p onto the direction: $p \rightarrow \langle d_p, p \rangle$, similarly for q, the direction is d_q. A sample of new coordinate is obtained: $S_{p,d_p} = (\langle d_p, p_1 \rangle, \ldots, \langle d_p, p_n \rangle)$ and similarly $S_{q,d_q} = (\langle d_q, q_1 \rangle, \ldots, \langle d_q, q_n \rangle)$. First step is to choose d_p and d_q for maximizing the correlation between vectors, such that:

$$\rho = \max_{d_p, d_q} Corr(S_p d_p, S_q d_q) \tag{1}$$

$$= \max_{d_p, d_q} \frac{\langle S_p d_p, S_q d_q \rangle}{\|S_p d_p\| \, \|S_q d_q\|} \tag{2}$$

Fig. 15. Overview of literature based on image–text cross-modal retrieval.

where ρ represents the equation to be maximized. Let E denotes the empirical expectation of function $f(p, q)$ and given by

$$E = \frac{1}{m}\sum_{i=1}^{m} f(p_i, q_i) \tag{3}$$

then ρ can be redefined as

$$\rho = \max_{d_p, d_q} \frac{E[\langle d_p, p\rangle \langle d_q, q\rangle]}{\sqrt{E[\langle d_p, p\rangle^2]E[\langle d_q, q\rangle^2]}} \tag{4}$$

$$= \max_{d_p, d_q} \frac{E[d_p' p q' d_q]}{\sqrt{E[d_p' p p' d_p]E[d_q' q q' d_q]}} \tag{5}$$

$$= \max_{d_p, d_q} \frac{d_p' E[p q'] d_q}{\sqrt{d_p' E[p p'] d_p d_q' E[q q'] d_q}} \tag{6}$$

Covariance matrix of (p, q) is defined as:

$$Cov(p, q) = E\left[\binom{p}{q}\binom{p}{q}'\right] = \begin{bmatrix} C_{pp} & C_{pq} \\ C_{qp} & C_{qq} \end{bmatrix} = C \tag{7}$$

Table 4
List of acronyms in alphabetical order.

Acronym	Definition	Acronym	Definition	Acronym	Definition
AADAH	Attention-Aware Deep Adversarial Hashing	DVSH-Q	DVSH variant without bitwise max-margin loss	SCM	Semantic Correlation Matching
ACMR	Adversarial Cross-Modal Retrieval	FSH	Fusion Similarity Hashing	*SCM_KL*	SCM with Kullback–Leibler divergence measure
AICDM	Annotation by Image-to-Concept Distribution Model	FSH-S	FSH with a simple fusion graph construction	*SCM_l1*	SCM with l1 distance measure
BoVW	Bag of Visual Words	GSS-SL	Generalized Semisupervised Structured Subspace learning	*SCM_l2*	SCM with l2 distance measure
BoW	Bag of Words	HOG	Histogram of Oriented Gradients	*SCM_NC*	SCM with normalized correlation measure
CCA	Canonical Correlation Analysis	HRL	Hybrid Representation Learning	*SCM_NC$_c$*	SCM with centered normalized correlation measure
CM	Correlation Matching	*HRL_C*	Hybrid Representation Learning with CNN features	SDCH	Semantic Deep Cross-modal Hashing
CM_l1	CM with l1 distance measure	*HRL_H*	Hybrid Representation Learning with handcrafted features	SIFT	Scale Invariant Feature Transformation
CM_l2	CM with l2 distance measure	IMH	Inter Media Hashing	SM	Semantic Matching
CM_NC	CM with normalized correlation measure	JFSSL	Joint Feature Selection and Subspace Learning	*SM_KL*	SM with Kullback–Leibler divergence measure
CM_NC$_c$	CM with centered normalized correlation measure	KCCA	Kernel Canonical Correlation Analysis	*SM_l1*	SM with l1 distance measure
CMOLRS	Cross-Modal Online Low-Rank Similarity	LBP	Local Binary Pattern	*SM_l2*	SM with l2 distance measure
CMSTH	Cross-Modal Self-Taught Hashing	LCMH	Linear Cross-Modal Hashing	*SM_NC*	SM with normalized correlation measure
CMTC	Cross-Modal Topic Correlation	LDA	Latent Dirichlet Allocation	*SM_NC$_c$*	SM with centered normalized correlation measure
CNN	Convolutional Neural Network	LSSH	Latent Semantic Sparse Hashing	SMFH	Supervised Matrix Factorization Hashing
Corr-AE	Correspondence Autoencoder	M3R	Multi-Modal Mutual topic Reinforcement modeling	SRLCH	Subspace Relation Learning for Cross-modal Hashing technique
CVH	Cross-View Hashing	MAP	Mean Average Precision	SVM	Support Vector Machine
DAML	Deep Adversarial Metric Learning	MDCR	Modality Dependent Cross-media Retrieval	TDH	Triplet based Deep Hashing
DAML_D	Deep Adversarial Metric Learning with deep features	MDSSL	Multiordered Discriminative Structured Subspace Learning	*TDH_C*	TDH with CNN-F features
DAML_S	Deep Adversarial Metric Learning with shallow features	MFDH	Multi-view Feature Discrete Hashing	*TDH_H*	TDH with handcrafted features
DCMH	Deep Cross-Modal Hashing	MHTN	Modal-adversarial Hybrid Transfer Network	UCH	Unsupervised Concatenation Hashing
DDL	Discriminative Dictionary Learning	MLRank	Multi-correlation Learning to Rank	*UCH_LLE*	UCH with Locally Linear Embedding
DLSH	Discrete Latent Semantic Hashing	MRR	Mean Reciprocal Rank	*UCH_LPP*	UCH with Locality Preserving Projection
DMSH	Deep Multi-level Semantic Hashing	MSFH	Multi-modal graph regularized Smooth matrix Factorization Hashing	ZSH	Zero-Shot Hashing
DVSH	Deep Visual Semantic Hashing	PR curve	Precision–Recall curve	ZSH1	ZSH (with complete data)
DVSH-B	DVSH variant without binarization	QCH	Quantized Correlation Hashing	ZSH2	ZSH (with zero shot data)
DVSH-H	DVSH variant without using the hashing networks	RCH	Robust Cross-view Hashing	ZSH3	ZSH (with semi-supervised zero shot data)
DVSH-I	DVSH variant by replacing the cosine max-margin loss	S3CA	Shared Semantic Space with Correlation Alignment	ZSH4	ZSH (with semi-supervised zero shot data and different label spaces)

Total covariance matrix C is a block matrix where $C_{pq} = C'_{qp}$ are between-sets covariance matrices and C_{pp}, C_{qq} are within-sets covariance matrices, although Eq. (7) represents the covariance matrix in zero-mean case only. Now ρ can be redefined as:

$$\rho = \max_{d_p, d_q} \frac{d'_p C_{pq} d_q}{\sqrt{d'_p C_{pp} d_p d'_q C_{qq} d_q}} \tag{8}$$

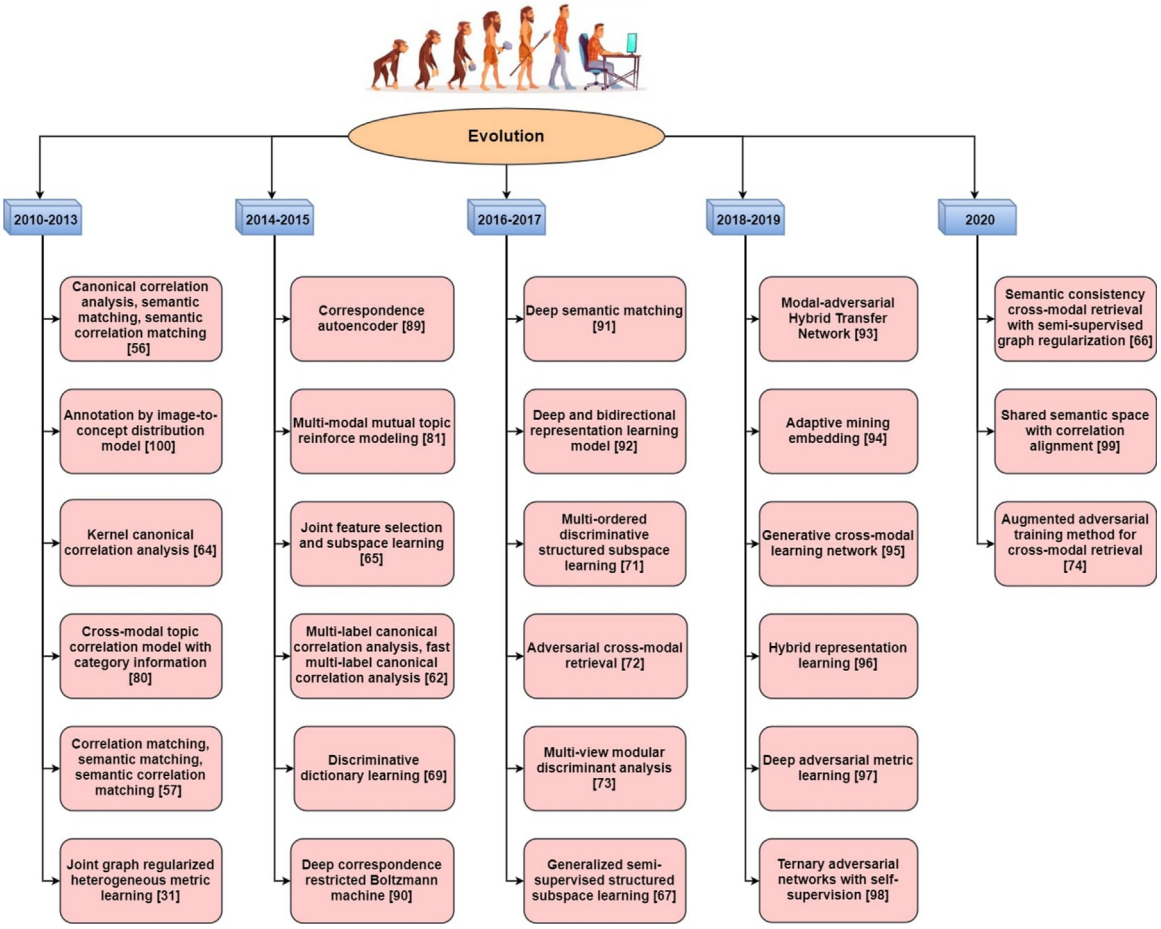

Fig. 16. Real-valued representation learning methods' research evolution.

The maximum of ρ w.r.t. d_p and d_q is the maximum canonical correlation.

Authors in [56] have proposed the use of CCA, Semantic Matching (SM), and Semantic Correlation Matching (SCM) in cross-modal document retrieval task. Two hypotheses have been investigated in this: (a) Benefit to explicitly modeling the correlation between two elements, and (b) This modeling is more useful in feature spaces having higher levels of abstraction. Images are represented by Scale Invariant Feature Transformation (SIFT) features and text using Latent Dirichlet Allocation (LDA). The motive is to retrieve images that closely match the text query and to retrieve text which closely matches the image query. A new Wikipedia dataset has been composed for the experimentation. The cross-modal framework proved to outperform the state-of-art cross-modal retrieval methods and even the novel image retrieval systems on the uni-modal retrieval tasks. A mathematical formulation is introduced in [57] which associates cross-modal retrieval systems' design with isomorphic feature spaces for diverse modalities. Two hypotheses are inspected related to the principal characteristics of these feature spaces: (1) low-level cross-modal correlations should be accounted for, and (2) space should allow semantic abstraction. So, three novel solutions to cross-modal retrieval problem are then obtained from these hypotheses are CM, SM and SCM. CM is an unsupervised approach that models cross-modal correlations, SM is a supervised method that relies on semantic representation and SCM is the combination of both of them.

In [58], a cross-modal retrieval framework has been presented which outputs a ranked list of semantically relevant text from a separate text corpus (having no related images) when queried using an image and vice versa. For these two tasks, a novel Structural SVM based unified formulation has been proposed. Two representations considered for both image and text modality are: (a) uni-modal probability distributions over topics learned using LDA, and (b) explicit learning multi-modal correlations using CCA. The work done in [41] is an extension of [58]. A new loss function based on normalized correlation is introduced in this which is found to be better than the previous two loss functions. Along with this, the proposed method is compared with other baseline methods, extensive analysis of training, and run-time efficiency. Comparison based on two new evaluation metrics and recent image and text features is also incorporated in the new work. [59] has proposed a cross-modal technique for extracting semantic relationship between classes using annotated images. Firstly, both visual features and text are projected onto a latent space using CCA, and then the probabilistic interpretation of CCA is utilized for calculating the representative distribution of the latent variable for each class. Two measures are obtained based on the representative distributions: (1) semantic relation between classes, and (2) abstraction level of each class.

Classic CCA method has few drawbacks [54]: (1) It is able to compute only the linear correlation between two sets of variables, however, the relationship may be non-linear in most of the real-world implementations; (2) It is able to operate only on two modalities; (3) If it is applied on a supervised problem then it wastes the information available in the form of labels because it is an unsupervised technique, and (4) Intra-modal semantic consistency is an important factor to improve retrieval accuracy but CCA fails to capture this [60]. To handle the drawbacks of classic CCA, several variants of this method are introduced such as

Generalized CCA (GCCA), Kernel CCA (KCCA), Locality Preserving CCA (LPCCA), and Deep CCA (DCCA) to name a few. CCA extension techniques seek to construct a correlation that maximizes non-linear projection. In [61], authors have introduced a new dataset containing images, text (paragraph), and hyperlinks. This dataset is named as *WIKI-CMR* and it is composed of Wikipedia articles. It consists of total of 74961 documents including images, textual paragraphs, and hyperlinks. Documents are classified into 11 diverse semantic classes. CCA and KCCA cross-modal retrieval techniques have been applied to the dataset. An Improved CCA (ICCA) technique has been proposed in [60] to control the limitations of traditional 2-view CCA. For improvement in intra-modal semantic consistency, two effective semantic features are proposed which are based on text features. Traditional 2-view CCA has been expanded to 4-view CCA and it is embedded into an escalating framework to reduce the over-fitting. The framework combines training of linear projection and non-linear hidden layers to make sure that fine representations of input raw data are learned at the output of the network. A similarity metric is also presented for improving distance measure which is inspired by large scale similarity learning. In [62], an extension of the CCA approach has been introduced, named multi-label CCA (ml-CCA). It learns the shared subspaces by taking care of high-level semantic information in the formation of multi-label annotations. This approach utilizes the multi-label information for generating correspondences instead of relying on explicit pairing among different modalities like CCA. A fast ml-CCA technique is also presented in this which has the capability of handling huge datasets.

An unsupervised learning framework based on KCCA is proposed which identifies the relation between image annotation by humans and the corresponding importance of things and their layout in the scene [63]. This uncovered relation is utilized in increasing the accuracy of search results as per queries. A novel approach for image retrieval and auto-tagging has been introduced in [64] which utilizes the object importance information provided by keyword tag list. It is an unsupervised approach based on KCCA which finds the relationship between image tagging by humans and the corresponding importance of objects and their outline in the scene. As the KCCA technique is non-parametric, so it scales poorly with the training set size and has trouble with huge real-world datasets [2]. To handle KCCA drawbacks and to provide an alternative, Deep CCA (DCCA) has been proposed. It tackles the scalability issue and leads to better correlated representation space.

Graph regularization based methods. Cross-modal retrieval typically includes two fundamental issues: (a) Relevance estimation; and (b) Coupled feature selection. In [65], authors are dealing with both the issues. To deal with the first issue, multi-modal data is mapped to a common subspace to measure the similarity among modalities. Projection matrices are learned for this mapping and l_{21}-norm penalties are imposed on them separately to deal with the second issue, which selects appropriate and discriminative features from diverse feature spaces at the same time. Further, a multi-modal graph regularization term is applied to the projected data to preserve intra and inter modality similarity relationships. An iterative algorithm is introduced for solving the joint learning issue along with its convergence analysis. The excessive experimentation on three popular datasets proved the proposed technique to outperform the state-of-art techniques.

To overcome the semantic and heterogeneity gap between modalities, the potential correlation of diverse modalities need to be considered. Also, the semantic information of class labels required to be utilized for reducing the semantic gap among different modalities as well as realizing the inter-dependence and interoperability of divergent modalities. So, authors in [52] have proposed a cross-modal retrieval framework which is based on graph regularization and modality dependence, fully utilizing the correlation between modalities. After considering the semantic and feature correlation, projection matrices are learned separately for Image-to-Text and Text-to-Image retrievals. Then the internal arrangement of original feature space is utilized to construct an adjoining graph having semantic information constraints which enables the diverse labels of miscellaneous modality data to get closer to respective semantic information. The whole process can be visualized in Fig. 18. The objective function for I2T and T2I tasks are defined in Eqs. (9) and (10) respectively.

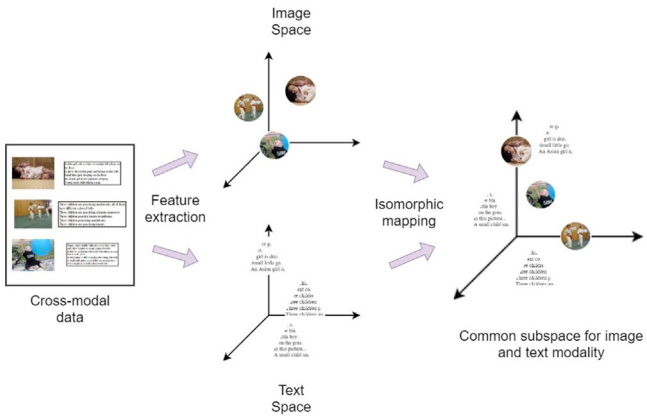

Fig. 17. General representation of subspace learning process.

$$F(U_1, V_1) = \lambda \left\| U_1^T X - V_1^T Y \right\|_F^2 + (1-\lambda) \left\| U_1^T X - S \right\|_F^2 + \alpha tr(U_1 X^T L_1 X U_1^T - S^T L_1 S) + \beta_1 \|U_1\|_F^2 + \beta_2 \|V_1\|_F^2 \quad (9)$$

$$F(U_2, V_2) = \lambda \left\| U_2^T X - V_2^T Y \right\|_F^2 + (1-\lambda) \left\| V_2^T Y - S \right\|_F^2 + \alpha tr(V_2 Y^T L_2 Y V_2^T - S^T L_2 S) + \beta_1 \|U_2\|_F^2 + \beta_2 \|V_2\|_F^2 \quad (10)$$

where U_1, U_2 and V_1, V_2 represent the image and text projection matrices in I2T and T2I respectively. S is the semantic matrix of image and text, X and Y represents the feature matrices of image and text respectively, λ, α, β_1 and β_2 are balance parameters. A semantic consistency cross-modal retrieval with semi-supervised graph regularization (SCCMR) method is introduced in [66] which ensures a globally optimal solution by merging prediction of labels and optimization of projection matrices to a unified architecture. Simultaneously, the method also considers nearest neighbors in potential image–text subspace and image–text with the same semantics using graph embedding. Discriminative features are captured from different modalities by applying l_{21}-norm constraint to projection matrices.

Inspired by the fact that unlabeled data can be composed easily and aid to exploit the correlation between modalities, [67] has proposed a novel framework generalized semi-supervised structured subspace learning (GSS-SL) for cross-modal retrieval. A label graph constraint is proposed for predicting appropriate concept labels for un-annotated data. For modeling correlation between modalities, GSS-SL utilizes the label space as a linkage after consideration of the fact that concept labels directly unveils the semantic information of multi-modal data. Specifically, a joint minimization formulation is created from the combination of the label-linked loss function, label graph constraint, and regularization for learning discriminative common subspace. Multiple

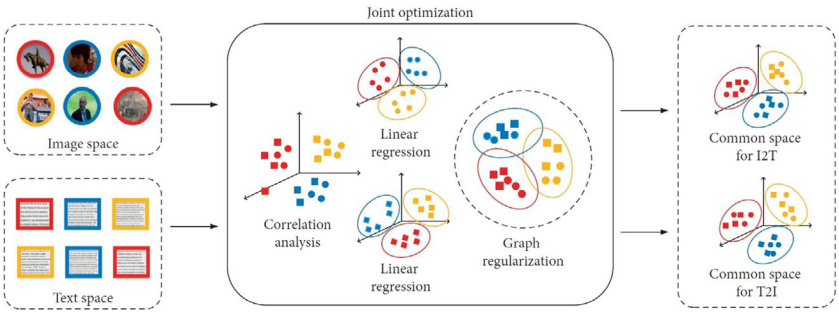

Fig. 18. Process of cross-modal retrieval framework followed in [52].

linear transformations are alternatively optimized by an effective optimization method for diverse modalities and updating of the class indicator matrices for un-annotated data is also performed.

Other subspace learning methods. A modality-dependent cross-media retrieval (MDCR) model has been proposed in [68] in which two couple of projections are learned for diverse cross-media retrieval tasks rather than one couple of projections. Two couple of mappings are learned to project text and images from original feature space into separate common latent subspaces by simultaneously optimizing the correlation between text and images and linear regression from one modal space to semantic space. A novel discriminative dictionary learning (DDL) approach amplified with common label alignment has been introduced in [69] for effective cross-modal retrieval. It increases the discriminative ability of intra-modality information from diverse concepts and relevance of inter-modality information in the same class. To handle the huge multi-modal web data, [70] has proposed a cluster-sensitive cross-modal correlation learning framework. A novel correlation subspace learning technique which learns a group of a cluster-sensitive sub-models is presented to better fit the content divergence of various modalities.

A Multi-ordered Discriminative Structured Subspace Learning (MDSSL) approach is proposed in [71]. This metric learning framework learns a discriminative structured subspace where data distribution is reserved for ensuring a required metric. An adversarial cross-modal retrieval method has been proposed in [72] which attempts to make an effective common subspace based on adversarial learning. To handle the problem of multi-view embedding from diverse visual hints and modalities, a unified solution is proposed for subspace learning techniques which makes use of Rayleigh quotient [73]. It is extendable for supervised learning, multiple views, and non-linear embedding. A multi-view modular discriminant analysis (MvMDA) approach is introduced for considering the view difference. After getting motivation from the fact that un-annotated data can be easily compiled and helps to utilize the correlations among diverse modalities, a novel generalized semi-supervised structured subspace learning (GSS-SL) approach is proposed in [67] for the task of cross-modal retrieval. For aligning diverse modality data by moving one source modality to another target modality, a cross-modal retrieval approach with augmented adversarial training is proposed in [74]. An augmented version of the conditional generative adversarial network is utilized for reserving the semantic meaning in the modality transfer process.

4.1.2. Statistical and probabilistic methods

Statistical methods include the Markov model (MM), Hidden Markov Model (HMM), Markov Random Field, and so forth. Probabilistic methods incorporate the use of probability and various probabilistic models. They are typically utilized to find out the probability of generating a particular modality result based on a given query modality. Scientific biomedical articles contain multi-modal information such as images and text. Considering the growth of the healthcare industry, important text and images keep on hiding under the inessential data which makes it hard to retrieve the relevant information. Biomedical articles often contain annotation markers or tags such as letters, stars, symbols, or arrows in figures which highlight the crucial area in the figure. These markers are also correlated with the image captions and text in the article. Identification of the markers becomes important to extract the ROIs from images. A novel technique has been proposed in [26] with the combination of rule-based and statistical image processing ways for localizing and annotating the medical image regions or ROIs. Moreover, a cross-modal image retrieval technique has been implemented on articles and it is based upon ROI identification and classification.

Automatic image annotation and retrieval framework based on probabilistic models have been proposed in [75] with an assumption that image regions can be explained using *blobs* (a kind of vocabulary). Blob is an acronym for Binary Large Object and it is a collection of binary data that is stored as a single unit in a database. Blobs are created from image features using clustering. To automatically annotate or retrieve images using a word as a query, the trained probabilistic model predicts the probability of producing a word with the help of image blobs. After experimentation, the proposed probabilistic model based on the cross-media relevance model is proved to be almost six times better than a model based on the word-blob co-occurrence model and two times better than a model derived from machine translation in terms of mean precision. An improvement of cross-media relevance model [75] is presented in [76] to automatically assign related keywords to un-annotated images based on images' train data. Images present in the training dataset are fragmented into parts and then these parts are represented using a blob. K-means algorithm is used for blobs' creation for clustering those image parts. Using this model, the probability for assigning a keyword into a blob is predicted and after annotation success, one image part is represented by a keyword. TF–IDF method is used for text document feature extraction and appropriate text documents are retrieved using images' automatic annotation information. Experimentation is performed on IAPR TC-12 and 500 Wikipedia web-pages (landscape related) dataset to show the usefulness of the proposed technique.

4.1.3. Rank based methods

These methods see the issue of cross-modal retrieval as a problem of learning to rank. Ranking of images and tags is suitable for efficient tag recommendation or image search. In [77], a new Multi-correlation Learning to Rank (MLRank) approach is proposed for image annotation which ranks the tags for images as per their relevance after considering semantic importance and visual similarity. Two cases are defined: (a) image-bias consistency; and (b) tag-bias consistency that is developed into an optimization problem for rank learning.

In [78], a ranking model has been optimized as a listwise ranking problem considering cross-modal retrieval process and a learning to rank with relational graph and pointwise constraint (LR2GP) technique has been proposed. Firstly, a discriminative ranking model is introduced that utilizes the relationship between a single modality for improvement in ranking performance and learning of an optimal embedding shared subspace. A pointwise constraint is proposed in low-dimension embedding space to make up for the real loss in the training phase. In the end, a dynamic interpolation algorithm is selected for dealing with the problem of fusion of loss function. A Cross-Modal Online Low-Rank Similarity function learning (CMOLRS) technique is proposed in [79] that learns a low-rank bilinear similarity measurement for the task of cross-modal retrieval. A fast-CMOLRS technique is also introduced which has less processing time than the former technique.

4.1.4. Topic models

Topic models are a kind of statistical model that finds the abstract topics which arise in a set of documents. A cross-modal topic correlation model has been introduced in [80] which jointly models the text and image modalities. A statistical correlation model is examined which is conditioned on category information. [81] proposed a novel supervised multi-modal mutual topic reinforcement modeling (M3R) technique that makes a joint cross-modal probabilistic graphical model for finding the mutually consistent semantic topics using required interaction between model factors.

A topic correlation model (TCM) is presented in [82] by mutual modeling of images and text modalities for cross-modal retrieval task. Images are represented by the bag-of-features model based on SIFT and text is represented by topic distribution learned from the latent topic model. These features are mapped into a common semantic space and statistical correlations are analyzed. These correlations are utilized for finding out the conditional probability of results in one modality while querying in another modality.

4.1.5. Machine learning and deep learning based methods

Machine learning (ML) refers to the capability of a machine to enhance its performance on the basis of previous outcomes. ML approaches allow systems to learn without being programmed explicitly. Deep learning mimics the way the human brain works for both feature extraction and classification as discussed in [83]. This section includes the works which are based on machine learning and deep learning. Summary of deep learning based cross-modal systems incorporating image and text have been presented separately in the Table 18. In [40], authors have proposed a novel technique of multi-modal Deep Belief Network for finding out the missing data in text or image modality. Also, the proposed model can be used for multi-modal data retrieval as well as annotation purpose. After experimentation on MIR Flickr data containing images and corresponding tags, the proposed model is found to be better than bi-modal data of images and text. Moreover, its performance outperforms the performance of Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) models. As the cross-modal data is heterogeneous in nature, so it is troublesome to compare directly. For making it comparable, authors in [30] have made use of deep learning by proposing a deep correlation mining technique. Various media features are trained in this technique and then fused together with the help of correlation between their trained features. Moreover, the Levenberg–Marquardt technique has been used for avoiding the local minima problem in deep learning. Experiments are performed on image–audio and image–text databases to validate the proposed solution. Authors have proposed a novel cross-modal retrieval technique based on similarity theory and deep learning [84]. They have utilized Local Binary Pattern (LBP) as an image descriptor and Deep Belief Network (DBN) as a deep learning algorithm.

In [85], a new Scalable Deep Multi-modal Learning (SDML) data retrieval method has been introduced. A common subspace is predefined to maximize between-class variation and minimize within-class variation. Then a network is trained for each modality separately such that n networks are obtained for n modalities. It is done to transform multi-modal data into the common predefined subspace for achieving multi-modal learning. The method is scalable to a number of modalities as it can train different modality-specific networks separately. It is the first proposed technique which is individually projecting data of different modalities into a predefined common subspace. Experimentation is performed on four benchmark datasets such as PKU XMedia, Wikipedia, NUS-WIDE, and MS-COCO dataset to validate the proposed technique. To solve the problem of image–text cross-modal retrieval, various novel models are introduced in [86] which are designed by correlating hidden representations of a pair of autoencoders. Minimizing correlation learning error enables the model to learn invisible representations by just utilizing the general information in diverse modalities. On the other hand, minimizing the representation learning error builds hidden representations good enough for reconstructing inputs of each modality. A specific parameter is set in the models to make a balance between two types of error generated by representation and correlation learning. Models are divided into two groups: (1) one contains three models that reconstruct both modalities and so named as multimodal reconstruction correspondence autoencoder, and (2) the second contains two models that reconstruct a single modality and so named as unimodal reconstruction correspondence autoencoder. Experimentation is performed on three popular datasets and the proposed technique is found to be better than two popular multimodal deep models and three CCA based models.

Supervised cross-modal retrieval techniques provide better accuracy than unsupervised techniques at the additional cost of data labeling or annotation. Lately, semi-supervised techniques are gaining popularity as they provide a better framework to balance the trade-off between annotation cost and retrieval accuracy. A novel deep semi-supervised framework is proposed in [87] to handle both annotated and un-annotated data. Firstly, an un-annotated part of training data is labeled using the label prediction component and then a common representation of both modalities is learned to perform cross-modal retrieval. The two modules of the network are trained in a sequential manner. After extensive experimentation on pascal, Wikipedia, and NUS-WIDE datasets, the proposed framework is found to be outperforming both supervised and semi-supervised existing methods. In [88], authors have introduced an image–text multi-modal neural language model which can be utilized for retrieving related images from complex sentence queries and vice versa. It has been presented here that text representations and image features can be jointly learned in the case of image–text modeling by training the models in conjunction using a convolutional network.

A novel correspondence autoencoder model is proposed in [89] which is designed by correlating hidden representations of two uni-modal autoencoders. For this model training, an optimal objective that minimizes the linear combination of representation learning errors for every mode and correlation learning error between the hidden representation of the modalities. A correspondence restricted Boltzmann machine (Corr-RBM) is proposed in [90] for mapping the original features of modality data into a low-dimensional common space where heterogeneous data can be compared. Two deep neural structures are made from corr-RBM as the chief building block for the cross-modal retrieval

process. Cross-modal retrieval is performed using CNN visual features with various classic approaches in [91]. A deep semantic matching (DSM) technique is also introduced for handling cross-modal retrieval w.r.t. samples labeled with one or multiple labels. In [92], authors have proposed a deep and bidirectional representation learning model (DBRLM) where images and text are represented by two separate convolutional based networks.

A novel modal-adversarial hybrid transfer network has been proposed in [93]. It realizes the knowledge transfer from the single-modal source domain to the cross-modal target domain and then learns the common cross-modal representation. The architecture is based on deep learning and is divided into two subnetworks: (a) Modal-sharing knowledge transfer subnetwork; and (b) Modal adversarial semantic learning subnetwork. A deep learning model has been introduced in [94], named, AdaMine (ADAptive MINing Embedding) for learning the common representation of recipe items incorporating recipe images and their recipe in textual form. In [95], authors have proposed a novel approach generative cross-modal learning network (GXN) which includes generative processes into the cross-modal feature embedding which will be useful in learning both global abstract features and local grounded features. A deep neural network based approach known as hybrid representation learning (HRL) is proposed for learning common representation for each modality [96].

A new deep adversarial metric learning (DAML) technique is introduced for cross-modal retrieval which maps annotated data pairs of diverse modalities non-linearly into a shared latent feature subspace [97]. The inter-concept difference is maximized and the intra-concept difference is minimized. Each data pair difference caught from modalities of the same class is also minimized. Motivated by zero-shot learning, [98] has presented a ternary adversarial network with self-supervision (TANSS) model. It includes three parallel sub-networks: (1) two semantic feature learning subnetworks which capture the intrinsic data structures of diverse modes and preserve their relationships using semantic features in shared semantic space; (2) a self-supervised semantic subnetwork that utilizes seen and unseen label word vectors to use them as guidance for supervising semantic feature learning and increases knowledge transfer to unseen labels; and (3) adversarial learning scheme is used for maximizing the correlation and consistency of semantic features among various modalities. This whole network facilitates effective iterative parameter optimization. In [99], a shared semantic space with correlation alignment (S3CA) is proposed for cross-modal data representation. It aligns the non-linear correlations of cross-modal data distribution in deep neural networks made for diversified data.

4.1.6. Other methods

This section includes the summary of those works which cannot be classified under any of the above classes. In [100], authors have proposed an Annotation by Image-to-Concept Distribution Model (AICDM) for image annotation using the links between visual features and human concepts from image-to-concept distribution. There is a rapid increase in the discussions regarding disaster and emergency management on social media these days. Flood event observation has a principal role in emergency management and the related videos and images are also uploaded and searched on the web while disasters. This data can be helpful in emergency management by using it in sensors. Inspired by this, authors in [25] are performing image retrieval enhancement in the field of floods and flood aids. Integration of image and text features is performed after extracting visual features from images using BoW and text features using TF–IDF and weirdness. After extensive experimentation on US FEMA and Facebook datasets, it has been demonstrated that the proposed method is enhancing the emergency management efficiency by showing improvement in image recognition with the incorporation of text features in it.

Images are ranked as per similarity of semantic features in the query by semantic example retrieval. So, in [38], the accuracy of semantic features is improved using cross-modal regularization which is based on associated text.

4.2. Binary representation learning or cross-modal hashing

In general, the word *hash* means *chop and mix* which consecutively means that the hashing function chops and mixes information to obtain hash results [101]. The idea of hashing was first introduced by *H. P. Luhn* in his 1953 article *A new method of recording and searching information* [102]. Entire information regarding the birth of hashing is presented in [103]. It is nearly impossible to achieve a completely even distribution. It can only be created by considering of structure of keys. For a random group of keys, it is impractical to generate an appropriate generic hash function as the keys are not known beforehand. Random uniform hash works best in this case. So, inspired by the need of using random access system having a huge capacity for business applications, *Peterson* gave an estimation for the amount of search needed for the exact location of a record in numerous storage systems including the sorted-file and index table method [104]. Then the term *hashing* was first used by *Morris* in his article [105] in 1968. Few general definitions in hashing are described below [101]:

- *Hashing function*: This function ($h(\cdot)$) is used to map the random size of data to a fixed interval $[0, p]$. Given a data having n data points i.e. $A = [a_1, a_2, \ldots, a_n] \in R^D$ (real coordinate space of dimension D) and a hashing function $h(\cdot)$, then $h(A) = [h(a_1), h(a_2), \ldots, h(a_n)] \in [0, p]$ are known as hashes or hash values of data points represented by A. Hashing function is practically utilized in a hash table data structure which is highly popular for quick data lookup.
- *Nearest neighbor (NN)*: It represents one or more data entities in $A = [a_1, a_2, \ldots, a_n] \in R^D$ which are nearest to the query point a_q.
- *Approximate nearest neighbor (ANN)*: It attempts to find a data point $a_x \in A$ which is an $\varepsilon-$approximate nearest neighbor of the query point a_q in that $\forall a_x \in A$, the distance between a_x and a satisfies the relation $d(a_x, a) \leq (1 + \varepsilon)d(a_q, a)$.

Cross-modal hashing techniques are effective in resolving the issue of large scale cross-modal retrieval because it combines the benefits of classic cross-modal retrieval and hashing. These techniques either rely on annotated training data or they lack semantic analysis [106]. For correlating diverse modalities, typical cross-modal hashing techniques learn a unified hash space. Then the search process is improved based on hash codes. Hashing methods are broadly classified into *Data-dependent* and *Data-independent* methods [107]. In data-dependent methods, an appropriate hash function is learned using the available training data, however, the hash function is generated using random mapping independent of the training data in data-independent methods. Hash function learning is categorized into two stages: (1) Dimensionality reduction; and (2) Quantization. Dimensionality reduction means mapping the information from the original space to a low-dimensional spatial representation. Quantization means a linear or non-linear transformation of actual features to binary segment the feature space for acquiring hash codes. The aim of hashing methods is to minimize the semantic gap among modalities as much as possible. A typical resolution for this issue can be learning of a uniform hash code to make it more consistent. Another resolution can be the minimization of the

Table 5
Comparison of hashing methods on the basis of various characteristics. T = Traditional hashing method and D = Deep learning based hashing method.

Characteristics	Type	Hashing method	Methods
Optimization	Relaxation	T	LCMH [108], QCH [109]
		D	TDH [110], SDCH [111]
	Discrete	T	DLSH [112], SRLCH [113]
		D	DVSH [114], DCMH [115]
	Alternative solution	T	MFDH [116], MSFH [117], SMFH [118]
		D	ZSH [119]
Hash function	Linear	T	UCH [120], LCMH [108], CMSTH [106], MSFH [117]
	Non-linear	T	SRLCH [113]
		D	DVSH [114]
Distance metric	Cosine	T	QCH [109]
		D	DVSH [114]
	Euclidean	T	LCMH [108], MSFH [117]
		D	ZSH [119]
	Hamming	T	DLSH [112], CMSTH [106]
		D	DCMH [115], TDH [110], AADAH [121]

Fig. 19. Evolution of research in cross-modal hashing.

coding distance and enhance its compactness. Hashing taxonomy followed in this survey is: (1) General hashing methods which are defined first; and (2) Deep learning based hashing methods which are defined later in a different subsection. General hashing methods include all the methods which do not incorporate deep learning. Fig. 19 presents an evolution of cross-modal hashing techniques.

Table 5 presents the comparison of hashing techniques on various characteristics such as optimization, time complexity, hash function, and distance metric utilized for similarity calculation. While optimizing the objective function, either the *relaxation* is given for easy optimization or not which we call *discrete* type. Relaxation of discrete hash codes may result in quantization loss and performance degradation [117]. Time complexity mentioned here is for the whole method execution where n is the number of training samples used in it. Hashing models can be categorized into linear and non-linear type [113]. The distance metric is the metric utilized in the inter or intra similarity among modalities' calculation.

4.2.1. General hashing methods

This section includes all the cross-modal retrieval works based on hashing technique and which does not incorporate a deep learning approach. In [120], authors have proposed an Unsupervised Concatenation Hashing (UCH) technique where Locally Linear Embedding and Locality Preserving Projection are introduced for reconstructing the manifold structure of original space in the hamming space. $l_{2,1}$-norm regularization is imposed on the projection matrices for exploiting the diverse characteristics of various modalities. The proposed technique has been compared with other hashing techniques such as CVH, IMH, RCH, FSH, and CCA [122] as well. CVH [123] is an extension of classic uni-modal spectral hashing [124] to multi-modal field. In IMH [125], learned binary codes conserve both inter and intra-media consistency. FSH [126] embeds the graph-based fusion similarity to a common hamming space. In RCH [127], common hamming space is learned in diverse modalities' binary codes are created as consistent as possible. Table 6 shows the comparison of these techniques when applied on Wikipedia and Pascal dataset. This comparison is based on MAP scores when images are retrieved from the text (*T2I*, the text is retrieved from image (*I2T*) and the average of both scores. Bold values in the table represent the highest MAP score in the respective task and hash code length.

In [106], authors have introduced Cross-Modal Self-Taught Hashing (CMSTH) technique for both cross-modal and uni-modal image retrieval. It can successfully catch the semantic correlation from un-annotated training data. Three steps are followed in the learning procedure: (1) Hierarchical Multi-Modal Topic Learning (HMMTL) is proposed for identifying multi-modal topics using semantic information; (2) Robust Matrix Factorization (RMF) is utilized for transferring the multi-modal topics to hash codes which form a unified hash space, and (3) in the end hash functions are learned for projecting the modalities to a unified hash space. A new cross-modal hashing technique is proposed in [108] to handle the method scalability issue in the training period. The time complexity of the technique varies linearly with training data size which allows scalable indexing for multi-media search over various modalities. Hash functions are learned accurately while considering inter and intra modality similarities. Experiments are performed on NUS-WIDE and Wikipedia dataset to prove the effectiveness of the method. The objective function utilized here for preservation of inter-similarity between modalities for the bi-modal case is defined as:

$$\begin{aligned} &\min_{B^{(1)},B^{(2)}} \left\| B^{(1)} - B^{(2)} \right\|_F^2; \\ &s.t., B^{(i)^T} e = 0, \\ &b^{(i)} \in \{-1, 1\}, \\ &B^{(i)^T} B^{(i)} = I_c, i = 1, 2; \end{aligned} \tag{11}$$

where $B^{(1)}$ and $B^{(2)}$ represents the data matrices of image and text modalities, e is $n \times 1$ vector having each entry equal to 1, $\|\cdot\|_F$ is a Frobenius norm, I_c is $c \times c$ identity matrix, B depicts final binary codes obtained, constraint $B^{(i)^T} e = 0$ needs each bit has same chance to be 1 or -1 and constraint $B^{(i)^T} B^{(i)} = I_c$ requires the bits of each modality to be acquired separately. Loss function term

Table 6
Comparison of benchmark techniques on the basis of MAP scores on Wikipedia and Pascal VOC dataset with different hash code lengths presented in [120].

Tasks	Methods	Length of hash codes							
		Wikipedia				Pascal VOC 2007			
		16	32	64	128	16	32	64	128
I2T	CVH [123]	0.1499	0.1408	0.1372	0.1323	0.1484	0.1187	0.1651	0.1411
	CCA [122]	0.1699	0.1519	0.1495	0.1472	0.1245	0.1267	0.123	0.1218
	IMH [125]	0.2022	0.2127	0.2164	0.2171	0.2087	0.2016	0.1873	0.1718
	RCH [127]	0.2102	0.2234	0.2397	0.2497	0.2633	0.3013	0.3209	0.333
	FSH [126]	0.2346	0.2491	0.2531	0.2573	0.289	0.3173	0.334	**0.3496**
	UCH_LPP [120]	0.242	0.2497	0.255	0.2576	0.2706	0.3074	0.3255	0.3277
	UCH_LLE [120]	**0.2429**	**0.2518**	**0.2578**	**0.2588**	**0.2905**	**0.3245**	**0.3345**	0.3396
T2I	CVH	0.1315	0.1171	0.108	0.1093	0.0931	0.0945	0.0978	0.0918
	CCA	0.1587	0.1392	0.1272	0.1211	0.1283	0.1362	0.1465	0.1553
	IMH	0.1648	0.1703	0.1737	0.172	0.1631	0.1558	0.1537	0.1464
	RCH	0.2171	0.2497	0.2825	0.2973	0.2145	0.2656	0.3275	0.3983
	FSH	0.2149	0.2241	0.2332	0.2368	0.2617	0.303	0.3216	0.3428
	UCH_LPP	0.2351	0.2518	0.2623	0.2689	0.3945	0.4877	0.5187	0.5321
	UCH_LLE	**0.2363**	**0.2567**	**0.2845**	**0.2993**	**0.4106**	**0.4913**	**0.5217**	**0.5343**
Average	CVH	0.1407	0.129	0.1226	0.1208	0.1208	0.1066	0.1315	0.1165
	CCA	0.1643	0.1456	0.1384	0.1341	0.1264	0.1315	0.1347	0.1386
	IMH	0.1835	0.1915	0.1951	0.1946	0.1859	0.1787	0.1705	0.1591
	RCH	0.2137	0.2365	0.2611	0.2735	0.2389	0.2834	0.3242	0.3657
	FSH	0.2248	0.2366	0.2431	0.247	0.2753	0.3102	0.3278	0.3462
	UCH_LPP	0.2385	0.2508	0.2586	0.2632	0.3326	0.3976	0.4221	0.4299
	UCH_LLE	**0.2396**	**0.2542**	**0.2712**	**0.2791**	**0.3506**	**0.4079**	**0.4281**	**0.437**

$\left\|B^{(1)} - B^{(2)}\right\|_F^2$ obtains the maximal consistency (or the minimal difference) on two object representations. Eq. (11) is extended for more than two modality case and the new general equation obtained is:

$$\begin{aligned} &\min_{B^{(i)}, i=1,\ldots,p} \sum_{i=1}^{p}\sum_{i<j}^{p} \left\|B^{(i)} - B^{(j)}\right\|_F^2; \\ &s.t., B^{(i)^T} e = 0, \\ &b^{(i)} \in \{-1, 1\}, \\ &B^{(i)^T} B^{(i)} = I_c, i = 1, \ldots, p, \end{aligned} \tag{12}$$

where p represents no. of diverse modalities and rest of the notations are same as Eq. (11).

The issue of cross-modal hashing is how to efficiently construct the correlation among diverse modality representations in the hash function learning process. Most of the traditional hashing techniques map the miscellaneous modality data to a joint abstraction space by linear projections similar to CCA. Due to this, these methods are unable to effectively reduce the semantic gap among modalities which has been proved to lead to better accuracy in information retrieval. So to tackle this issue, a Latent Semantic Sparse Hashing method has been proposed in [128]. This method executes the cross-modal similarity with the use of sparse coding, for capturing important images' structures, and matrix factorization, for learning latent concepts from the text. In [109], a quantized correlation hashing (QCH) technique is proposed which considers the quantization loss over different modalities and the relation among them simultaneously. The relation among diverse modalities that explains the similar object is established by maximizing the correlation between the hash codes across modes. The resultant objective function is converted to a uni-modal formulation which is then optimized using another process. Objective function is defined in Eq. (13). Suppose two modalities (x_i, y_i) are representing n object, where x_i^T depicts ith row of data matrix $X \in R^{n \times d_x}$ of one modal and y_i^T represents ith row of data matrix $Y \in R^{n \times d_y}$ of another modal. d_x and d_y are dimensions of the modalities. Similarity information between data points across domains is defined as: $S_{ij} = 1$ iff x_i and y_j are similar and 0 otherwise.

$$\begin{aligned} &min\ O(B_x, B_y, W_x, W_y) = (\|B_x - XW_x\|_F^2 + \\ &\left\|B_y - YW_y\right\|_F^2) - \alpha' \sum_{(i,j)} S_{ij} \\ &\left(x_i^T W_x W_y^T y_j - \sqrt{x_i^T W_x W_x^T x_i}\sqrt{y_j^T W_y W_y^T y_j}\right) \\ &s.t.\ W_x^T W_x = I_{c\times c} \\ &W_y^T W_y = I_{c\times c} \end{aligned} \tag{13}$$

where $B_x \in \{-1, 1\}^{n\times c}$ and $B_y \in \{-1, 1\}^{n\times c}$ are two kinds of binary codes with same code length c for each object. $W_x \in R^{d_x \times c}$ and $W_y \in R^{d_y \times c}$ depicts two projection matrices for two modalities, W_y^T means transpose of a matrix W_y and similarly for other matrices. α' represents control parameter for balancing quantization loss and cosine similarity constraint. For making W_x and W_y as orthogonal projections, constraints $W_x^T W_x = I_{c\times c}$ and $W_y^T W_y = I_{c\times c}$ are used.

Most of the classic hashing techniques either suffer from high training costs or fail to capture the diverse semantics of various modalities. In order to tackle this issue, [112] has presented an efficient Discrete Latent Semantic Hashing (DLSH) approach. Firstly, it learns the latent semantic representations of miscellaneous modalities and afterward, projects them into a common hamming space for supporting scalable cross-modal retrieval. This approach directly correlates the explicit semantic labels with binary codes, so it increases the discriminative ability of learned hashing codes. Unlike traditional hashing approaches, DLSH directly learns binary codes using an effective discrete hash optimization. The overall objective function of the DLSH approach for

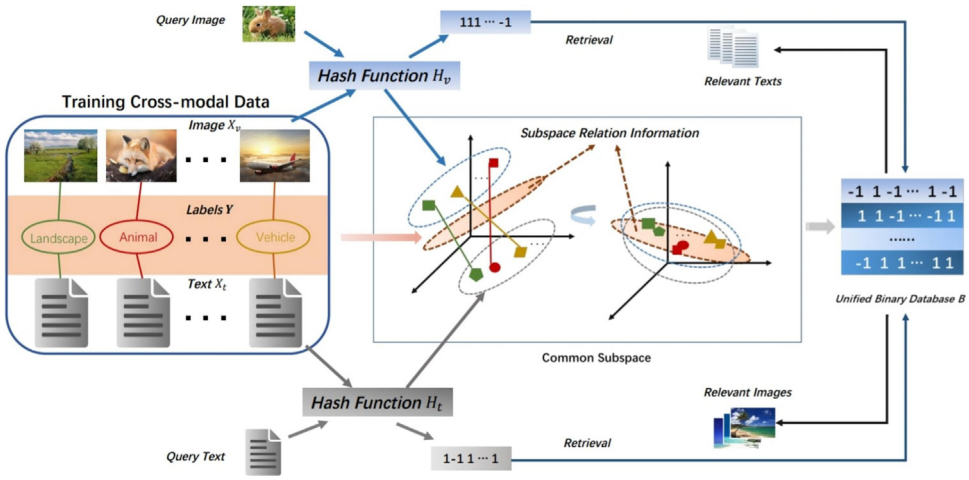

Fig. 20. Cross-modal hashing approach proposed in [113].

two modalities is given as:

$$\min_{U_i|_{i=1,2},A_i|_{i=1,2},W_i|_{i=1,2},Q} \sum_{i=1}^{2} \|\phi(X_i) - U_i A_i\|_F^2 + \beta \sum_{i=1}^{2} \|B - W_i A_i\|_F^2 + \delta \|B - QY\|_F^2 + \gamma \left(\sum_{i=1}^{2} \|U_i\|_F^2 + \sum_{i=1}^{2} \|W_i\|_F^2 + \|Q\|_F^2 \right) \quad s.t. B \in \{-1, 1\}^{L \times N} \tag{14}$$

where B is binary hash code matrix, $\|\cdot\|_F$ is the Frobenius norm of matrix, L is hash code length and N is no. of training instances, X_i denotes the original feature matrices of modalities, Q is semantic transfer matrix, $A_i \in R^{k \times N}$ is the latent semantic representation of modalities and k is its dimension, $U_i \in R^{m \times k}$ is basis matrix and m is no. of anchors, $W_i \in R^{L \times k}$ represents projection matrices for two sub-retrieval tasks, $\phi(X_i) \in R^{m \times N}$ is Gaussian kernel projection of image and text features, β and δ are penalty parameters and γ is regularization parameter for avoiding over-fitting.

In [113], authors have proposed a novel supervised Subspace Relation Learning for Cross-modal Hashing technique which utilizes the relation information of labels in semantic space for making similar data from diverse modalities nearer in the low-dimension hamming subspace. This technique preserves the discrete constraints, modality relations, and non-linear structures while admitting a closed-form binary code solution which increases the training accuracy. Both hash functions and unified binary codes are learned at the same time using an iterative alternative optimization algorithm. Using these hash functions and binary codes, multi-modal data can be effectively indexed and searched. The framework of the proposed SRLCH technique is shown in Fig. 20.

In [118], authors have proposed an approach of supervised matrix factorization hashing for using label information and effective cross-modal retrieval. This method is based on collective matrix factorization which considers both local geometric consistency in each mode and label consistency across several modalities. To resolve the issue of quantization loss which happens by relaxing discrete hash codes in the cross-modal retrieval process, [117] has proposed a multi-modal graph regularized smooth matrix factorization hashing which is an unsupervised technique. The aim of this technique is to learn unified hash codes for multi-media data in a common latent space where similarity of diverse modalities can be identified efficiently.

[116] utilizes multiple views for image and text representation to enhance feature information. A discrete hashing learning framework is proposed which employs complementary information among multiple views to make discriminative compact hash codes learning better. It performs classifier and subspace learning simultaneously for completing multiple searches at the same time.

4.2.2. Cross-modal hashing methods based on deep learning

Deep learning has become highly popular in recent years. Features extracted by deep learning methods have a powerful capability of expressing the data and they also have rich semantic information contained in them [106]. Thus, the multi-media information retrieval accuracy enhances significantly by combining hashing methods with deep learning. Various works incorporating cross-modal hashing methods based on deep learning have been introduced recently which are discussed in this section.

Capturing of spatial dependency of images and temporal dynamics of text is an important task in learning potential feature representations and cross-modal relations as it reduces the heterogeneity gap among modalities. So, a novel Deep Visual Semantic Hashing model has been introduced in [114]. It creates concise hash codes of textual sentences and images in a complete deep learning architecture that catches the essential cross-modal correspondences between natural language and visual data. DVSH model has a hybrid deep framework that comprises a visual semantic fusion network to learn joint embedding space of text and images, and two mode-specific hashing networks to learn hash functions for generating concise binary codes. The proposed framework efficiently unites cross-modal hashing and joint multi-modal embedding that is based on a new amalgamation of RNN over sentences, CNN over images, and a structures max-margin objective which combines everything together to facilitate the learning of similarity preserving and high-quality hash codes. Various cross-modal hashing techniques are based on hand-crafted features that may not attain a good accuracy value. A novel deep cross-modal hashing technique has been introduced in [115] by combining hash-code learning and feature learning into the same framework. From beginning to end, this framework consists of deep neural networks, one for each mode to do feature learning from starting.

A triplet based deep hashing network is proposed in [110]. Firstly, the triplet labels are utilized that explains the relative relationship among three instances as supervision for catching more common semantic correlations among cross-modal instances. For boosting the discriminative ability of hash codes,

a loss function is generated from intra-modal and inter-modal views. In the end, graph regularization is utilized for preserving the actual semantic similarity between hash codes in the hamming space. A deep adversarial hashing network has been proposed in [121] with attention mechanism for increasing the measurement of content similarities for particularly aiming at the informative pieces of multi-media. It has three modules: (a) feature learning module for getting feature representations; (b) attention module for creating attention mask; (c) hashing module for learning hash functions. A novel deep cross-modal hashing framework is proposed in [111] which combines hash codes and feature learning into the same network. It has considered both inter and intra modality correlation and a loss function with dual semantic supervision for hash learning.

In [119], a novel cross-modal zero-shot hashing method has been introduced which efficiently utilizes both labeled and unlabeled multi-modal data having separate label spaces. Zero-shot hashing learns a hashing model that is trained using only samples from seen classes, however, it has the capability of good generalization for unseen classes' samples. Typically, it utilizes the class attributes to seek a semantic embedding space for transferring knowledge from seen classes to unseen classes. So, it may perform poorly in the case of less labeled data. In [129], authors have proposed a multi-level semantic supervision generating method after exploring the label relevance, and a deep hashing framework is introduced for multi-label image–text cross-modal retrieval. It can capture the binary similarity as well as the complex multi-label semantic structure of data in diverse forms at the same time.

5. Benchmark datasets

With the advent of huge multi-modal data generation, cross-modal retrieval has become a crucial and interesting problem. Researchers have composed diverse multi-modal datasets for evaluating the proposed cross-modal techniques. Fig. 21 presents the evolution of the datasets in recent years. Summary of prominent multi-modal datasets is given in Table 7 which includes dataset name, mode, total concepts, dataset size, image representation, text representation, related article, and data source. Fig. 22 presents a graph of the total number of categories in the datasets. Information regarding prominent benchmark datasets is given in the following points. After going through all the references related to cross-modal retrieval used in this survey, approximately used frequencies of popular datasets have been found and are represented in the form of a bar chart in Fig. 23.

1. *NUS-WIDE*[1] [130]: This is a real-world web image dataset composed by *Lab for Media Search* in the National University of Singapore. It consists of: (a) 2,69,648 images and associated tags from Flickr with 5018 unique tags, (b) Ground-truth for 81 concepts; and (c) low-level image features of six types, comprising 144-D color correlogram, 128-D wavelet texture, 64-D color histogram, 500-D BoW based on SIFT descriptions, 73-D edge direction histogram and 225-D block-wise color moments. Fig. 24 shows two examples (angelfish and autumn class) from the dataset with the image and the associated tags.
2. *IAPR TC-12*[2] [131]: This dataset is also known as ImageCLEF 2006. It has been created for CLEF (Cross-Language Evaluation Forum) cross-language image retrieval task. It is composed of 20,000 images taken from a private photographic image collection and associated captions are in

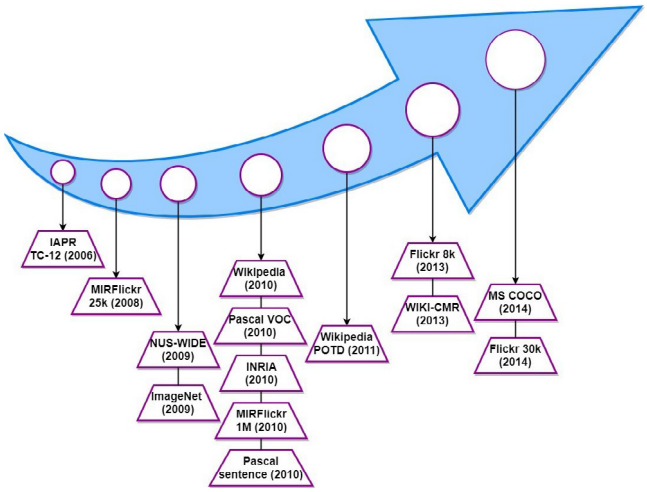

Fig. 21. Evolution of benchmark datasets.

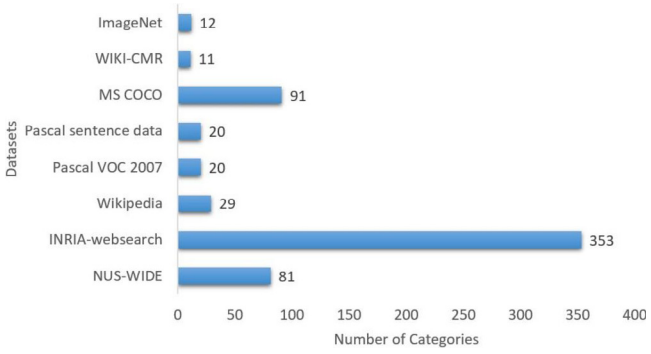

Fig. 22. A chart displaying the total number of categories in the popular datasets.

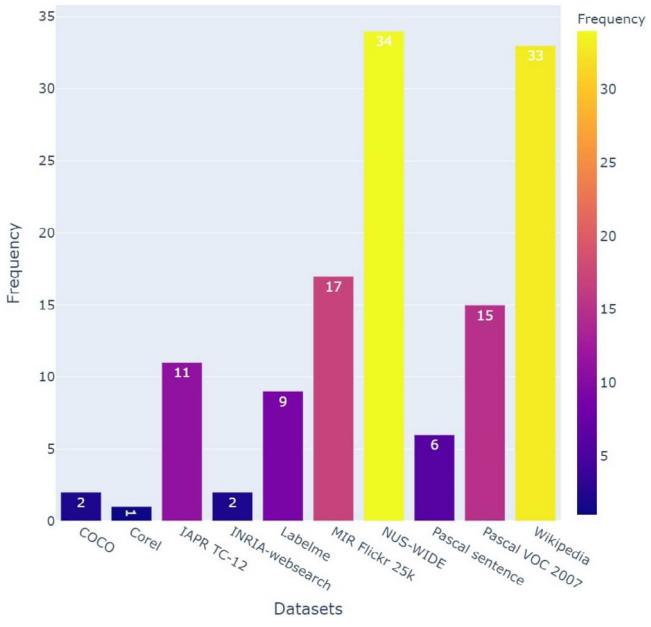

Fig. 23. Approximate used frequencies of prominent datasets in the references on cross-modal retrieval.

[1] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html

[2] https://www.imageclef.org/photodata

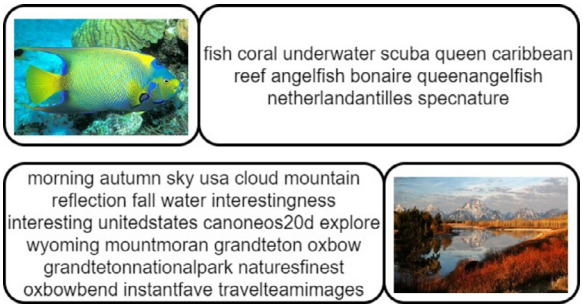

Fig. 24. Two examples from NUS-WIDE dataset in which an image is associated with numerous related tags.

three different languages such as English, Spanish, and German. This benchmark has been established from an initiative started by Technical Committee 12 (TC-12) of the International Association of Pattern Recognition (IAPR). The idea behind this dataset creation was to use it for evaluating the efficiency of both visual and text-based retrieval techniques.

3. *Wikipedia*[3] [56]: It consists of a document corpus with associated text and image pairs. It has been designed from Wikipedia's featured articles which are complemented by one or more images from Wikipedia Commons, providing a pair of desirable variety. Each article is classified into one of 29 concepts by Wikipedia and the concepts are assigned to both image and text modules of the article. The researchers have considered the top 10 highly populated concepts as some of the concepts are rare. The final corpus consists of 2866 documents. These are image–text pairs that have been assigned a class from the vocabulary of 10 semantic classes.
4. *PASCAL VOC 2007*[4] [132]: This dataset has been taken from the PASCAL (pattern analysis, statistical modeling, and computational learning) VOC (Visual Object Challenge) challenge. The dataset provided in this challenge is being utilized by researchers for the evaluation of the proposed cross-modal techniques. PASCAL VOC 2007 dataset has been widely used by the research community. It contains annotated consumer pictures composed from *Flickr*[5] (photo and video sharing website). The dataset consists of a total of 9963 images and 24,640 annotated objects which have been categorized into 20 different classes with four main concepts. The images consist of varied viewing conditions such as lightning, pose, and others. Annotators took guidance from annotation guidelines[6] for appropriately annotating each image in the ground-truth [133]. The entities mentioned in the annotation are class, bounding box, view, truncated, and difficult.
5. *MIR FLickr 25k and 1M*[7] [134,135]: The dataset is available in 2 sizes: 25k and 1M. The images have been collected from Flickr for the research purpose related to image content and image tags. Moreover, tags and EXIF (Exchangeable image file format) image metadata has also been extracted and made publicly available. Image tags have been presented in two forms: (a) raw form in which they are obtained from users; and, (b) in the processed form where raw tags have been cleaned by Flickr (e.g. removal of spaces and special characters). In MIR Flickr 25k data, images have been manually annotated. Each image has an average of 8.94 tags. So, there are 1386 tags that are associated with at least 20 images. Images are split into 15,000 training and 10,000 testing images. MIR Flickr 1M data is an extension of MIR Flickr 25k. Images have not been annotated manually, unlike original 25k data. Images are represented using MPEG-7 edge histogram and homogeneous texture descriptors and color descriptors.
6. *INRIA-Websearch* [136]: This dataset consists of 71,478 images resulted from a web search engine for 353 miscellaneous search queries. Top-ranked images have been chosen from this search along with their corresponding metadata and ground-truth annotations. For each searched query, the dataset comprises the initial textual query, top-ranked images, and an annotation file. More than 200 images have been retrieved for 80% of queries. Annotation file consists of manual labels for image relevance to the query and other related metadata such as web page URL, image URL, page title, image's alternate text, 10 words before the image on a web page, and 10 words after. Images have been scaled to fit in a 150 $\times$ 150 pixel square, however, preserving the original aspect ratio.
7. *Flickr 8k and 30k*[8] [137,138]: Flickr 30k is an extension of the Flickr 8k dataset. Both datasets have been created from the Flickr website. Flickr 8k contains 8092 images and its main focus is on people or animals (mainly dogs) carrying out some action. Images have been collected from six different Flickr groups manually and annotated using multiple captions in the form of sentences by selected workers from the US. Flickr 30k contains 31,783 images of everyday scenes, activities, and events. Images are associated with 1,58,915 captions which have been attained via crowd-sourcing. The approach followed to collect this data is the same as followed by [137].
8. *PASCAL sentence data*[9] [139]: The images for this dataset have been collected from PASCAL VOC 2008 challenge [132]. Data consists of 1000 images selected from around 6000 images of PASCAL VOC 2008 training data. Images have been categorized into 20 categories depending upon the objects that appear in them and few images are present in multiple classes. Fifty random images have been chosen from each class to compose the dataset. Each image is annotated with five different captions in the form of sentences.
9. *MS-COCO*[10] [140]: Microsoft Common Objects in COntext (MS COCO) dataset has been composed of the pictures of daily scenes consisting of general objects in their usual environment. The objects are labeled using per-instance segmentation to help in precise object localization. The dataset consists of total 3,28,000 images with 25,00,000 labeled instances. The objects chosen for the dataset are from 91 diverse categories. The annotation pipeline has been divided into three prominent exercises: (1) labeling concepts which are present in the image, (2) locating and marking all instances of labeled concepts; and (3) segmentation of each object instance.
10. *WIKI-CMR* [61]: This dataset has been collected from Wikipedia articles which contain images, paragraphs and hyperlinks. Authors mainly focused on the areas: geography, people, nature, culture and history for dataset collection. It consists of total 74,961 documents categorized into

3 http://www.svcl.ucsd.edu/projects/crossmodal/
4 http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html
5 https://www.flickr.com/
6 http://host.robots.ox.ac.uk/pascal/VOC/voc2007/guidelines.html
7 http://press.liacs.nl/mirflickr/
8 http://shannon.cs.illinois.edu/DenotationGraph/
9 http://vision.cs.uiuc.edu/pascal-sentences/
10 http://cocodataset.org/

11 diverse concepts. Each of the document includes one paragraph, one associated image (or no image), a category label and hyperlinks. Images are represented using eight types of features including dense SIFT, Gist, PHOG, LBP and other features. Text is represented using TF–IDF.

6. Comparative analysis

In this section, prominent evaluation metrics used for cross-modal retrieval method performance analysis are defined. Afterward, comparisons of various cross-modal retrieval methods when applied on diverse datasets are presented on the basis of MAP score.

6.1. Evaluation metrics

For image and text modality, two cross-modal retrievals are considered: (a) image to text retrieval (I2T), means retrieving text related to the query image; and (2) text to image retrieval (T2I), retrieving images that match with the textual query [1]. Precisely in the testing phase, given a text or an image query, the aim of the cross-modal method is to search and retrieve the images or text that closely matches the query modality respectively. A retrieved outcome is considered to be relevant if it belongs to the same concept as the query modality. Two typical factors considered while quantitative performance evaluation are: (1) class relevance evaluation between query and outcome; (2) examining cross-modal relevance for image–text pairs. The first factor tells about the ability to learn diverse cross-modal latent representations while the second factor says about the capability of learning correlated latent concepts [81]. The metrics related to the above two factors are as follows:

1. *Precision, recall and PR curve*: *Precision* is defined as the ratio of TP to $TP+FP$, where TP is the number of outcomes which are similar to query and $TP + FP$ is the number of total retrieved outcomes. It is useful in measuring the probability of success for an information retrieval system. On the other hand, *Recall* is defined as the ratio of TP to $TP + FN$, where TP is the same as explained above and $TP + FN$ is the total number of relevant outcomes in the repository. It is useful in measuring the percentage of retrieved relevant results for an information retrieval system [76,84]. Refer to the Table 8 for a complete understanding of the definition of precision and recall. Precision and recall can be expressed as (Eqs. (15), (16)):
$$prec = \frac{TP}{TP + FP} \tag{15}$$
$$rec = \frac{TP}{TP + FN} \tag{16}$$
where *prec* represents precision, *rec* is recall, TP indicates true positive, FP is false positive and FN represents false negative.
Most of the works [52,143–145] have used the precision–recall curve to visualize the performance of their algorithm. The curve indicates the precision value at different recall levels. Authors in [146] have also used precision curve for performance visualization. It indicates the change in precision with respect to the number of retrieved results.
2. *F-measure*: It is a typical metric utilized for evaluating the performance of information retrieval systems [84]. After considering the effects of both precision and recall, F-measure can be defined mathematically as Eq. (17):
$$F = \frac{(\theta^2 + 1) \times prec \times rec}{\theta^2 \times (prec + rec)} \tag{17}$$
here θ has been used for adjusting the weighted proportion of both recall (rec) and precision (prec). If θ becomes 1 then F-measure can be redefined as F_1 (Eq. (18)):
$$F_1 = \frac{2 \times prec \times rec}{prec + rec} \tag{18}$$
Here, F_1 is the perfect combination of recall and precision. More the value of F_1, more better is the algorithm.
3. *MAP*: Mean Average Precision (MAP) is the most popular metric used for evaluating the performance of a cross-modal retrieval algorithm. It measures whether the retrieved result belongs to the same class as the query data (relevant) or not (irrelevant) [81]. It is the average of average precision calculated over all the queries. Given a query (an image or a text) and a group of its corresponding O retrieved outcomes, average precision is defined as (Eq. (19)):
$$AP = \frac{1}{R} \sum_{o=1}^{O} P(o) rel(o) \tag{19}$$
where R is the number of relevant outcomes in the retrieved outcomes, $P(o)$ is the precision of top o retrieved outcomes, if the oth retrieved outcome is relevant then $rel(o) = 1$ and otherwise 0. Now, MAP can be defined as (Eq. (20)):
$$MAP = \frac{1}{Q} \sum_{q=1}^{Q} AP \tag{20}$$
where Q is the total number of queries. A large MAP value signifies the betterment of the cross-modal algorithm when applied on a particular dataset.
4. *Percentage*: MAP metric only considers the factors that whether the outcome is relevant to query or not. For more precise evaluation, all the retrieved outcomes are ranked as per correlation. Typically, a query text (or image) is considered to be successful in retrieving results if its corresponding ground-truth image (or text) appears in the first a percent of the ranked list of retrieved outcomes. *Percentage* is the ratio of correctly retrieved query outcomes among all the query outcomes. Authors in [81,84,142] have utilized this metric for algorithm evaluation and have chosen the value of a as 0.2 or 20%.
5. *MRR*: Mean Reciprocal Rank (MRR) is another performance evaluation metric similar to percentage. It has been applied in [81,84] for method evaluation regarding the position of the corresponding ground-truth outcome paired with the query. It is mathematically expressed as (Eq. (21)):
$$MRR = \frac{1}{|O|} \sum_{n=0}^{|O|} \frac{1}{rank_n} \tag{21}$$
where $|O|$ is the number of query outcomes, $rank_n$ indicates the position of corresponding unique ground-truth paired with nth query in the retrieved set.

6.2. Comparison of results using diverse techniques

This section presents the comparison of various cross-modal retrieval techniques on primary datasets. Techniques are compared on the basis of the MAP score as it is the most popular and widely used evaluation metric. Three MAP scores are considered here which are *I2T* (when image queries related text), *T2I* (when text queries matched images), and the average of these (*I2T* and

Table 7
Summary of prominent image–text multi-modal datasets.

Sr. No.	Dataset	Year	Mode	Total concepts	Total images/ text	Image representation	Text representation
1	IAPR TC-12 [131]	2006	Image/ caption	Diverse	20,000/ 60,000	–	–
2	MIRFlickr 25k [134]	2008	Image/ tags	Diverse	25,000/ 2,23,500	–	–
3	NUS-WIDE [130]	2009	Image/ tags	81	2,69,648/ 5,018 unique tags	Color correlogram, wavelet texture, color histogram, BoW based on SIFT descriptions, edge direction histogram and block-wise color moments	Tag occurrence feature
4	ImageNet[a] [141]	2009	Images/ synsets	12 subtrees	32,00,000/ 5,247	SIFT	–
5	Wikipedia [56]	2010	Image/ text	29 (10 major)	2,866/ 2,866	SIFT	LDA
6	Pascal VOC 2007 [132]	2010	Image/ tags	20	9,963/ 24,640	–	–
7	MIRFlickr 1M [135]	2010	Image/ tags	Diverse	10,00,000/ 89,40,000	MPEG-7 edge histogram and homogeneous texture descriptors, color descriptor	Flickr user tags, EXIF metadata
8	INRIA-websearch [136]	2010	Image/ labels	353	71478/ -	–	–
9	Pascal sentence data [139]	2010	Image/ sentences	20	1000/ 5000	–	–
10	Wikipedia POTD [142]	2011	Images/ paragraphs	NA	1987/ 1987	SIFT	Text tokenization using rainbow
11	Flickr 8k [137]	2013	Image/ captions or sentences	Diverse	8092/ 40460	–	–
12	WIKI-CMR [61]	2013	Images/ paragraphs/ hyperlinks	11	38,804/ 74,961	SIFT, gist, PHOG, LBP, self similarity, spatial pyramid method	TF–IDF
13	Flickr 30k [138]	2014	Image/ captions or sentences	Diverse	31,783/ 1,58,915	–	–
14	MS COCO [140]	2014	Images/ labels	91	3,28,000/ 25,00,000	–	–

[a] http://www.image-net.org

Table 8
Table for better understanding of precision and recall.

	Relevant	Irrelevant	Total
Retrieved	True Positive (TP)	False Positive (FP)	Predicted Positive
Not retrieved	False Negative (FN)	True Negative (TN)	Predicted Negative
Total	Actual Positive	Actual Negative	$TP + FP + TN + FN$

T2I) two values. Table 9 shows the MAP scores on NUS-WIDE dataset. The blank spaces in the tables indicate that there is no value provided for that particular hash code length. The bold value in each of the hash code columns represents the highest value in that column. Fig. 25 presents a chart of average MAP scores for Table 9. It is evident from Table 9 and Fig. 25 that the performance of SDCH [111] method is the best in both I2T and T2I tasks, however, MFDH [116] shows best performance on 128 hash code length. Tables 10 and 11 presents the MAP scores on Wikipedia and MIRFlickr 25k dataset respectively. For Table 10, the best results are shown by MFDH [116] technique in both I2T and T2I tasks. ZSH1 [119] method shows the best performance on MIRFlickr 25k dataset for 128 hash code and otherwise SDCH [111] performs the best. Table 12 shows the comparison of various cross-modal hashing techniques based on deep learning on the IAPR TC-12 dataset. SDCH [111] method has the highest map score in both I2T and T2I tasks on all hash code lengths except 128. On length 128, DVSH-B [114] method shows the highest performance for both tasks. Average MAP results shown in Tables 10, 11 and 12 can be visualized in Figs. 26, 27 and 28 for Wikipedia, MIRFlickr and IAPR TC-12 datasets respectively.

Tables 13 and 14 show the comparison of various real-valued learning techniques based on MAP score on Wikipedia and NUS-WIDE dataset respectively. Four types of methods are included:

Table 9
Comparison of prominent hashing techniques on the basis of MAP scores on NUS-WIDE dataset with different hash code lengths.

Methods	Length of hash codes											
	I2T				T2I				Average			
	16	32	64	128	16	32	64	128	16	32	64	128
IMH [125]	0.2056	0.2145	0.2317		0.2381	0.2533	0.2613		0.22185	0.2339	0.2465	
LSSH [128]	0.4933	0.5006	0.5069	0.5084	0.625	0.6578	0.6823	0.6913	0.55915	0.5792	0.5946	0.59985
QCH [109]	0.5395	0.5489			0.5568	0.5741			0.54815	0.5615		
CMSTH [106]	0.5032	0.5073	0.527	0.5439	0.4761	0.4965	0.5088	0.5243	0.48965	0.5019	0.5179	0.5341
FSH-S [126]	0.4996	0.461	0.4556		0.4776	0.446	0.4423		0.4886	0.4535	0.44895	
FSH [126]	0.5059	0.5063	0.5171		0.479	0.481	0.4965		0.49245	0.49365	0.5068	
SMFH [118]	0.4553	0.4623	0.4658	0.468	0.5033	0.5056	0.5065	0.5079	0.4793	0.48395	0.48615	0.48795
MFDH [116]	0.646	0.6714	0.7014	**0.7121**	0.7811	0.8285	0.8653	**0.8824**	0.71355	0.74995	0.78335	**0.79725**
DLSH [112]	0.5127	0.516	0.5179		0.5203	0.5234	0.5284		0.5165	0.5197	0.52315	
DCMH [115]	0.5903	0.6031	0.6093		0.6389	0.6511	0.6571		0.6146	0.6271	0.6332	
AADAH [121]	0.6403	0.6294	0.652		0.6789	0.6975	0.7039		0.6596	0.66345	0.67795	
TDH_H [110]	0.6393	0.6626	0.6754		0.6647	0.6758	0.6803		0.652	0.6692	0.67785	
TDH_C [110]	0.6393	0.6626	0.6754		0.6647	0.6758	0.6803		0.652	0.6692	0.67785	
ZSH1 [119]	0.6411	0.6434	0.6457	0.6468	0.6755	0.6763	0.6789	0.6796	0.6583	0.65985	0.6623	0.6632
ZSH2 [119]	0.5982	0.6017	0.6033	0.6059	0.6286	0.6297	0.6325	0.6339	0.6134	0.6157	0.6179	0.6199
ZSH3 [119]	0.1733	0.1756	0.1771	0.1783	0.1721	0.1736	0.1743	0.1748	0.1727	0.1746	0.1757	0.17655
ZSH4 [119]	0.1481	0.1492	0.1511	0.1519	0.1437	0.1453	0.1475	0.1498	0.1459	0.14725	0.1493	0.15085
SDCH [111]	**0.813**	**0.834**	**0.841**		**0.823**	**0.857**	**0.868**		**0.818**	**0.8455**	**0.8545**	

Table 10
Comparison of prominent hashing techniques on the basis of MAP scores on Wikipedia dataset with different hash code lengths.

Methods	Length of hash codes											
	I2T				T2I				Average			
	16	32	64	128	16	32	64	128	16	32	64	128
LSSH [128]	0.233	0.234	0.2387	0.234	0.5571	0.5743	0.571	0.5577	0.39505	0.40415	0.40485	0.39585
QCH [109]	0.2343	0.2477			0.3034	0.317			0.26885	0.28235		
CMSTH [106]	0.3155	0.3293	0.3313	0.3375	0.3562	0.37	0.3825	0.3878	0.33585	0.34965	0.3569	0.36265
SMFH [118]	0.2572	0.2759	0.2863	0.2913	0.5784	0.604	0.6163	0.6219	0.4178	0.43995	0.4513	0.4566
MFDH [116]	**0.3548**	**0.3763**	**0.3878**	**0.3954**	**0.8318**	**0.8458**	**0.8568**	**0.8666**	**0.5933**	**0.61105**	**0.6223**	**0.631**
DLSH [112]	0.2838	0.3429	0.352		0.6764	0.7478	0.749		0.4801	0.54535	0.5505	
ZSH1 [119]	0.2998	0.3017	0.3035	0.3063	0.3016	0.3025	0.3044	0.3061	0.3007	0.3021	0.30395	0.3062
ZSH2 [119]	0.2543	0.2551	0.2576	0.2581	0.2526	0.2541	0.2563	0.2587	0.25345	0.2546	0.25695	0.2584
ZSH3 [119]	0.1214	0.1233	0.1247	0.1251	0.1178	0.1196	0.1218	0.1232	0.1196	0.12145	0.12325	0.12415
ZSH4 [119]	0.0982	0.0997	0.1012	0.1019	0.0936	0.0949	0.0971	0.0995	0.0959	0.0973	0.09915	0.1007

Table 11
Comparison of prominent hashing techniques on the basis of MAP scores on MIRFlickr 25k dataset with different hash code lengths.

Methods	Length of hash codes											
	I2T				T2I				Average			
	16	32	64	128	16	32	64	128	16	32	64	128
FSH-S [126]	0.609	0.5969	0.593		0.6036	0.5944	0.5923		0.6063	0.59565	0.59265	
FSH [126]	0.5968	0.6189	0.6195		0.5924	0.6128	0.6091		0.5946	0.61585	0.6143	
MFDH [116]	0.6836	0.6939	0.7066	0.723	0.7408	0.7506	0.7602	0.7797	0.7122	0.72225	0.7334	0.75135
DLSH [112]	0.6379	0.648	0.6603		0.6764	0.6777	0.685		0.65715	0.66285	0.67265	
DCMH [115]	0.741	0.7465	0.7485		0.7827	0.79	0.7932		0.76185	0.76825	0.77085	
AADAH [121]	0.7563	0.7719	0.772		0.7922	0.8062	0.8074		0.77425	0.78905	0.7897	
TDH_H [110]	0.711	0.7228	0.7289		0.7422	0.75	0.7548		0.7266	0.7364	0.74185	
TDH_C [110]	0.711	0.7228	0.7289		0.7422	0.75	0.7548		0.7266	0.7364	0.74185	
DMSH [129]	0.726	0.737	0.75		0.755	0.763	0.775		0.7405	0.75	0.7625	
ZSH1 [119]	0.7812	0.7831	0.7862	**0.7874**	0.7964	0.7989	0.8025	**0.8037**	0.7888	0.791	0.79435	**0.79555**
ZSH2 [119]	0.7302	0.7334	0.7351	0.7363	0.7092	0.7113	0.7132	0.7148	0.7197	0.72235	0.72415	0.72555
ZSH3 [119]	0.2126	0.2135	0.2141	0.2147	0.2016	0.2023	0.2027	0.2031	0.2071	0.2079	0.2084	0.2089
ZSH4 [119]	0.1873	0.1899	0.1917	0.1926	0.1795	0.1807	0.1816	0.1822	0.1834	0.1853	0.18665	0.1874
SDCH [111]	**0.845**	**0.866**	**0.873**		**0.831**	**0.856**	**0.863**		**0.838**	**0.861**	**0.868**	

(1) deep learning based; (2) subspace learning methods; (3) topic models; and (4) rank-based methods. Map score in bold font represents the highest value in that particular column and italic font represents the highest value in a particular method type.

7. Discussion

Cross-modal information retrieval is a burdensome task because of the semantic gap among modalities. Due to which different modalities cannot be compared directly to each other. To handle this issue, researchers have introduced several techniques for multi-modal data representation in the past few years. Table 18 presents a summary of recent literature for state-of-the-art techniques used for image–text cross-modal retrieval. It is divided into three parts: the first part contains works incorporating real-valued representation learning, the second includes binary representation learning works and the third is devoted to works based on deep learning. The table describes the cross-modal method, image and text feature extractors, the dataset used, method type, evaluation metric, and references.

Table 12
Comparison of prominent deep learning based hashing techniques on the basis of MAP scores on IAPR TC-12 dataset with different hash code lengths.

Methods	Length of hash codes											
	I2T				T2I				Average			
	16	32	64	128	16	32	64	128	16	32	64	128
DVSH [114]	0.5696	0.6321	0.6964	0.7236	0.6037	0.6395	0.6806	0.6751	0.58665	0.6358	0.6885	0.69935
DVSH-B [114]	0.626	0.6761	0.7359	**0.7554**	0.6285	0.6728	0.6922	**0.6756**	0.62725	0.67445	0.71405	**0.7155**
DVSH-Q [114]	0.5385	0.6113	0.6869	0.7097	0.5684	0.6153	0.6618	0.6693	0.55345	0.6133	0.67435	0.6895
DVSH-I [114]	0.4792	0.5035	0.5583	0.589	0.4903	0.5496	0.589	0.6012	0.48475	0.52655	0.57365	0.5951
DVSH-H [114]	0.4575	0.4975	0.5493	0.569	0.4396	0.4853	0.5185	0.5337	0.44855	0.4914	0.5339	0.55135
DCMH [115]	0.4526	0.4732	0.4844		0.5185	0.5378	0.5468		0.48555	0.5055	0.5156	
AADAH [121]	0.5293	0.5283	0.5439		0.5358	0.5565	0.5648		0.53255	0.5424	0.55435	
SDCH [111]	**0.726**	**0.787**	**0.803**		**0.704**	**0.783**	**0.797**		**0.715**	**0.785**	**0.8**	

Table 13
Performance comparison of prominent real-valued learning methods on the basis of MAP score on Wikipedia dataset.

Method type	Methods	MAP score		
		I2T	T2I	Average
Deep learning based methods	LBP+DBN [84]	0.2576	0.2761	0.2669
	Deep semi-supervised framework [87]	0.436	0.341	0.388
	MHTN [93]	0.514	0.444	0.479
	HRL_H [96]	**0.672**	**0.686**	**0.679**
	HRL_C [96]	0.647	0.666	0.656
	DAML_S [97]	0.356	0.267	0.322
	DAML_D [97]	0.559	0.481	0.52
	S3CA [99]	0.551	0.485	0.518
	Corr-AE [89]	0.326	0.361	0.344
	Corr-Cross-AE [89]	0.336	0.341	0.338
	Corr-Full-AE [89]	0.335	0.368	0.352
Subspace learning methods	JFSSL [65]	0.3063	0.2275	0.2669
	CM [56]	0.249	0.196	0.223
	SM [56]	0.225	0.223	0.224
	SCM [56]	0.277	0.226	0.252
	CM_l1 [57]	0.193	0.234	0.214
	CM_l2 [57]	0.199	0.243	0.221
	CM_NC [57]	0.288	0.239	0.263
	CM_NC$_c$ [57]	0.287	0.239	0.263
	SM_l1 [57]	0.22	0.274	0.247
	SM_l2 [57]	0.205	0.276	0.241
	SM_NC [57]	0.301	0.276	0.289
	SM_NC$_c$ [57]	0.352	0.272	0.312
	SM_KL [57]	0.206	0.274	0.24
	SCM_l1 [57]	0.334	0.273	0.304
	SCM_l2 [57]	0.315	0.267	0.291
	SCM_NC [57]	0.371	0.279	0.325
	SCM_NC_c [57]	0.382	0.281	0.332
	SCM_KL [57]	0.311	0.27	0.291
	MDCR [68]	0.287	0.225	0.256
	SCCMR [66]	*0.431*	*0.403*	*0.417*
	MDSSL [71]	0.3517	0.2851	0.3184
Topic models	CMTC [80]	*0.293*	0.232	*0.266*
	M3R [81]	0.2298	*0.2677*	0.2488
Rank based methods	CMOLRS [79]	*0.454*	*0.414*	*0.434*
	fast CMOLRS [79]	0.451	0.411	0.431

The data-dependent hashing methods can be categorized into supervised, unsupervised, and semi-supervised as per utilization of data supervision information. Supervised methods usually obtain better search accuracy than the other two methods because of the utilization of semantic label information. Unsupervised methods are appropriate for small scale and data-distributed retrieval, however, semi-supervised methods perform better in case of less label information. Table 15 shows the comparison of these three types of methods. Deep learning plays a vital role in hash learning, feature extraction, and retrieval performance in a hashing method. Usually, the deep learning based hashing method performs better than the general hashing method as it is data-dependent and its performance depends upon a substantial increase in data scale. So, deep hashing methods usually perform better in case of a colossal amount of multi-modal data but with higher hardware cost. Besides, the black box feature extraction attributes of deep learning may lead to the exclusion of vital information from the original data. Moreover, the optimization process of deep learning requires plenty of manual fine-tuning [107]. The refinement and potent of feature extraction part of the deep hashing method must be considered in future works. Table 16 presents a comparison of general and deep learning based hashing methods in the cross-modal retrieval field. As the hash retrieval is a type of statistical task, label information plays an important role in it without particular method consideration. In the case of un-annotated or incomplete labeled data, impulsively following retrieval performance under a supervised situation may lead to poor algorithm performance. So, consideration of algorithm performance in case of diverse data labeling degrees is required in the future.

Table 14
Performance comparison of prominent real-valued learning methods on the basis of MAP score on NUS-WIDE dataset.

Method type	Methods	MAP score		
		I2T	T2I	Average
Deep learning based methods	LBP+DBN [84]	0.3373	0.4221	0.3797
	Deep semi-supervised framework [87]	0.556	0.422	0.489
	MHTN [93]	0.52	0.534	0.527
	HRL_H [96]	0.446	0.476	0.461
	HRL_C [96]	**0.603**	**0.599**	**0.601**
	DAML_S [97]	0.531	0.539	0.535
	DAML_D [97]	0.512	0.534	0.523
	Corr-AE [89]	0.319	0.375	0.347
	Corr-Cross-AE [89]	0.349	0.348	0.349
	Corr-Full-AE [89]	0.331	0.379	0.355
Subspace learning methods	JFSSL [65]	0.4035	0.3747	0.3891
	SCCMR [66]	0.434	0.386	0.41
	DDL [69]	0.4498	0.498	0.4739
	MDSSL [71]	0.5218	0.4079	0.4649
	ACMR [72]	*0.544*	*0.538*	*0.541*
	GSS-SL [67]	0.5364	0.404	0.4702
Topic model	M3R [81]	*0.2445*	*0.3044*	*0.2742*
Rank based methods	CMOLRS [79]	*0.415*	0.34	0.378
	fast CMOLRS [79]	0.414	*0.348*	*0.381*

Table 15
Comparison of hashing techniques in diverse supervision modes.

Mode	Label use	Data process	Hash learning	Retrieval performance	Performance in huge data
Supervised	Yes	Complex	Complex	Good	Good
Unsupervised	No	Simple	Simple	Fair	Poor
Semi-supervised	Partly	Simple	Complex	Average	Fair

Table 16
Comparison of general and deep learning based hashing methods.

Hashing method	Generality	Modeling complexity	Retrieval performance	Parameter scale	Hardware cost
General	Poor	Complex	Fair	Small	Small
Deep learning based	Good	Simple	Good	Large	Large

Table 17
Open issues in cross-modal retrieval.

Type	Open issue	Current state
Algorithm level	1. Appropriate adoption of diverse modality feature descriptors	1. Descriptors provided with benchmark datasets are chosen mostly
	2. Need of a hybrid of soft and hard computing approaches	2. Either soft or hard computing approach is utilized
	3. Need of a scalable algorithm	3. Algorithms have restrictions of data size, modalities and application areas
	4. Need of a reproducible cross-modal retrieval method	4. Most methods are applicable in a particular application area
	5. Cross-modal retrieval implementation in big data, cloud, and IoT environments	5. Rarely applied
	6. More utilization of semi-supervised cross-modal retrieval techniques	6. Less used
Data level	7. Lack of huge datasets incorporating diverse modalities	7. Most existing benchmark datasets are old and consist of only image and text modality
	8. Requirement of proper and exact labeling of images	8. Poor and noisy annotations
	9. Diversity in data composition area	9. Datasets majorly composed from common social media websites

8. Open issues

The motive of cross-modal learning is to prepare a model to which one type of modality is inserted as a query to retrieve the results in another modality. For this process, the collected data has to be arranged in a manner so that retrieval can happen in less time as well as the results must be accurate and semantically relate to the queried modality data. Researchers have proposed miscellaneous algorithms for making cross-modal retrieval task more effective, however, there are few open issues which still need to be considered in the future to make the retrieval process much better. These issues are discussed in a Table 17 which can act as future research directions. The table is categorized into two parts: (1) Algorithm level; and (2) Data level. Former discusses

Table 18
Summary of works done in image–text cross-modal retrieval.

SR.	TECHNIQUE	IMAGE REP.	TEXT REP.	DATA	TYPE	METRIC	REF.
Real-valued representation techniques							
1	Linkage of each image feature with text feature	Bag-of-Words (BoW)	CiCui system [147], TF–IDF and weirdness [148]	US FEMA flood data, Facebook pages' and groups' data related to floods	–	MAP	[25]
2	Structural SVM, SR, CSR (using CCA)	BoW of dense SIFT features	Probability distribution	UIUC Pascal Sentence dataset, IAPR TC-12 benchmark and SBU-Captioned Photo dataset	–	BLEU score, rouge score	[58]
3	Markov Random Field (MRF) and Hidden Markov Model (HMM)	Image moments, gray level co-occurrence matrix (GLCM) moments, auto-correlation coefficients (AC), edge frequency (EF), Gabor filter descriptor, Tamura descriptor, color edge directional descriptor (CEDD), fuzzy color texture histogram (FCTH) descriptor and combined texture feature	Bag-of-Keywords	Thoracic CT scan data of nine distinct concepts containing 842 ROIs (created)	Supervised	Precision, recall and their curve, ten-fold cross validation accuracy, classification accuracy	[26]
4	Cluster sensitive cross-modal correlation learning framework	Wavelet feature, 3 level spacial max-pooling, GIST, dense SIFT with sparse coding, PHOG and color histogram	TF–IDF and Latent Dirichlet allocation(LDA)	Image Clef and Wikipedia [149] dataset	Semi-supervised	MAP	[70]
5	AICDM	Scalable color descriptor, color layout descriptor, homogeneous texture descriptor, edge histogram, grid color moment and gabor wavelet moment	–	ESP, Pascal VOC 2007, Web image	–	PR curve	[100]
6	Probabilistic model of automatic image annotation	Blobs to represent image regions	TF–IDF	IAPR TC-12 and 500 Wikipedia web-pages dataset	–	Precision, recall, F-measure	[76]
7	Joint feature selection and subspace learning	Gist, SIFT	LDA	Pascal, Wikipedia and NUS-WIDE dataset	–	MAP, PR curve	[65]
8	Local Group based Consistent Feature Learning (LGCFL)	GIST, HoG	Word frequency feature, latent dirichlet allocation model with 10 dimensions	LabelMe, Wikipedia, Pascal VOC2007, NUS-WIDE	Supervised	MAP, PR curve	[150]

(*continued on next page*)

the issues related to the cross-modal algorithm and later presents the open issues related to the multi-modal data considered in the algorithm. The current situation corresponding to the open issue is also described separately in the table.

1. *Noisy and restricted annotations*: A large amount of multi-modal data is created by people on various websites such as YouTube, Facebook, and Flickr to name a few. This data on the web is not properly organized and annotations

Table 18 (*continued*).

SR.	TECHNIQUE	IMAGE REP.	TEXT REP.	DATA	TYPE	METRIC	REF.
9	KCCA based approach	Gist, color histogram, BoVW	Word frequency, relative tag rank, absolute tag rank	Pascal VOC 2007, labelme,	Unsupervised	Normalized discounted cumulative gain	[64]
10	Structural SVM based unified framework	SIFT, BoVW, LDA	BoW, LDA	IAPR TC-12 Benchmark, UIUC Pascal Sentence, SBU-Captioned Photo	–	BLEU, precision, recall, median rank, MAP	[41]
11	Cross mOdal Similarity Learning algorithm with Active Queries (COSLAQ)	SIFT, GIST	latent Dirichlet allocation model	Wikipedia , Pascal VOC2007, NUSWIDE-1.5K, LabelMe	Supervised	MAP	[42]
12	CM, SM, SCM	SIFT	LDA	Wikipedia	Unsupervised	MAP, PR curve	[56]
13	Graph regularization and modality dependence (GRMD)	CNN	LDA	INRIA-websearch, Pascal sentence, Wikipedia 2010	–	MAP, PR curve	[52]
14	CCA, KCCA	SIFT, gist, PHOG, LBP, self similarity, spatial pyramid method	TF–IDF	WIKI-CMR	–	Precision	[61]
15	Improved CCA	–	–	NUS-WIDE, Pascal sentence, Wikipedia	–	MAP	[60]
16	Unsupervised KCCA based framework	Gist, HSV color histogram, SIFT	Words frequency, relative tag rank, absolute tag rank	Labelme, Pascal VOC	Unsupervised	Normalized Discounted Cumulative Gain (NDCG)	[63]
17	MLRank	Gist, color histogram, color auto-correlation, edge direction histogram, wavelet texture, block-wise color moments	–	Corel 5k, NUS-WIDE, IAPR TC12	Semi-supervised	Precision, recall, F1 score, MAP, N+ (no. of keywords with non-zero recall value)	[77]
18	CM, SM, SCM	SIFT	LDA	TVGraz, Wikipedia	Supervised and unsupervised	MAP, PR curve	[57]
19	CCA and its probabilistic interpretation	RGB-SIFT	Binary features	MIRFlickr 1M	–	Precision	[59]
20	Regularizer of image semantics	SIFT	LDA	TVGraz, Wikipedia, Pascal sentence dataset	–	MAP, PR curve	[38]
21	Modality-dependent cross-media retrieval (MDCR) model	CNN visual features	LDA	Wikipedia, Pascal sentence, INRIA-websearch	Supervised	MAP	[68]

(*continued on next page*)

related to it are also noisy and restricted. Annotations provide the required semantic information to understand the particular modality data and labeling a huge data manually is almost impossible. In [157], authors have used the combination of noisy and cleanly annotated images for robust image representations. One technique for combining noisy and clean data is to train a network with noisy data and then fine-tune it using clean data. However, this technique is not suitable for the proper usage of clean data. The proposed method represents the technique of using clean annotations for a reduction in noise in a large dataset and fine-tuning of the network with both clean and reduced noise data. The method consists of a multi-task network that learns to clean noisy annotations together with efficient classification of images. Extensive experimentation is performed on the Open Images dataset to show the efficiency of the proposed technique.

Table 18 (*continued*).

SR.	TECHNIQUE	IMAGE REP.	TEXT REP.	DATA	TYPE	METRIC	REF.
22	Semantic consistency cross-modal retrieval (SCCMR)	CNN, VGG	LDA, BoW	Wikipedia, Pascal sentence, NUS-WIDE-10k, INRIA-websearch	Semi-supervised	MAP, PR curve	[66]
Binary-valued or cross-modal hashing techniques							
23	Unsupervised Concatenation Hashing (UCH)	Gist	Word frequency count	Pascal, UCI handwritten digit, Wikipedia	Unsupervised	MAP	[120]
24	Cross-modal self-taught hashing (CMSTH)	SIFT, HoG, GIST	TF–IDF	Wikipedia, NUS-WIDE	Unsupervised	MAP	[106]
25	Linear cross-modal hashing	SIFT	LDA	NUS-WIDE, Wikipedia	–	MAP, recall	[108]
26	Latent semantic sparse hashing	Sparse coding	Matrix factorization	Labelme, NUS-WIDE, Wikipedia	–	MAP, PR curve	[128]
27	Quantized correlation hashing	SIFT, BoW	LDA, tag vector	58W-CIFAR, NUS-WIDE, Wikipedia	Supervised	MAP, precision	[109]
28	Discrete Latent Semantic Hashing	SIFT, gist, edge histogram	Topic vectors, index vector of selected tags, feature vector derived from PCA, binary tagging vector	Labelme, MIRFlickr 25k, NUS-WIDE, Wikipedia	Supervised	MAP, PR curve	[112]
29	Subspace Relation Learning for Cross-modal Hashing	SIFT, gist	LDA, tag occurrence feature vector	ImageNet, Labelme, MIRFlickr 25k, NUS-WIDE, UCI handwritten digit data, Wikipedia	Supervised	MAP, precision	[113]
30	Deep Visual Semantic Hashing model	Deep *fc*7 features	BoW vector	IAPR TC-12, MS COCO	–	MAP, precision	[114]
31	Deep cross-modal hashing	Gist, bag-of-visual-words (BOVW)	BoW vector	IAPR TC-12, MIRFlickr 25k, NUS-WIDE	–	MAP, PR curve	[115]
32	Triplet-based deep hashing network	SIFT	BoW	MIRFlickr 25k, NUS-WIDE	Supervised	MAP, PR curve	[110]
33	Attention-Aware Deep Adversarial Hashing	–	BoW	IAPR TC-12, NUS-WIDE, MIRFlickr 25k	–	MAP, PR curve	[121]
34	Supervised matrix factorization hashing	SIFT	Topic vector, BoW	NUS-WIDE, Wikipedia	Supervised	MAP, precision, PR curve	[118]
35	Semantic deep cross-modal hashing	–	BoW	IAPR TC-12, MIRFlickr, NUS-WIDE	Supervised	MAP, precision curve, PR curve	[111]
36	Zero-shot hashing	BoVW, SIFT, gist	LDA, BoW	MIRFlickr, NUS-WIDE, Wikipedia	Semi-supervised	MAP	[119]

(*continued on next page*)

2. *Need of a hybrid approach for designing cross-modal system*: Soft computing techniques have been used extensively these days to solve real-life problems and they show good results in data representation [158–160]. Although, various authors have applied these techniques for cross-modal system design, however, they are still in their inception stage and need to be explored more. Moreover, researchers either use a soft-computing or algorithmic approach. Both have their own limitations and strengths. So, there is a need for a hybrid approach to link heterogeneous data of various modalities.
3. *Lack of large scale multi-modal datasets*: Researchers are designing various algorithms these days for cross-modal retrieval and annotation. However, there is a lack of huge

Table 18 (*continued*).

SR.	TECHNIQUE	IMAGE REP.	TEXT REP.	DATA	TYPE	METRIC	REF.
37	Deep multi-level semantic hashing	–	BoW	MIRFlickr 25k	Supervised	MAP, PR curve	[129]
38	Cycle-Consistent Deep Generative Hashing (CYC-DGH)	CNN	LDA	Microsoft COCO, IAPR TC-12, wiki	–	MAP, precision curve, PR curve	[151]
39	Multi-modal graph regularized smooth matrix factorization hashing	SIFT, BoW, edge histogram	LDA, tag vector feature vectors	MIRFlickr, NUS-WIDE, Wikipedia	Unsupervised	MAP, PR curve	[117]
40	Multi-view feature discrete hashing	SIFT, histogram feature, BoVW	Word vector, mean vector, covariance matrix, feature histogram	MIRFlickr, MMED, NUS-WIDE, Wikipedia	Supervised	MAP, PR curve	[116]
Cross-modal methods based on deep learning							
41	Multi-modal Deep Belief Network (DBN)	Image specific DBN which used Gaussian Restricted Boltzmann Machines (RBM)	Text specific DBN which used Replicated Softmax model	MIR Flickr Data	Unsupervised	MAP	[40]
42	Levenberg–Marquardt deep canonical correlation analysis (LMDCCA)	Deep neural network	Deep neural network	Wikipedia Articles data	–	Precision recovery curve	[30]
43	Cross-media multiple deep network	GIST, Pyramid Histogram of Words (PHoW), MPEG-7, SIFT, color correlogram, color histogram, wavelet texture, edge direction histogram, block-wise color moments	BoW	NUSWIDE-10k, Wikipedia, Pascal Sentences	–	MAP, PR curve	[152]
44	Deep canonical correlation analysis (DCCA)	Color histogram, color correlogram, edge direction histogram, wavelet texture, block-wise color moments, SIFT, GIST, MPEG-7	BoW	Wikipedia, Pascal, NUS-WIDE10k	Supervised	MAP	[153]
45	Deep coupled metric learning (DCML)	SIFT, BoVW, GIST, color histogram	LDA model	Wikipedia, Pascal VOC 2007, NUS-WIDE	–	Precision, MAP, ROC and CMC curve	[154]
46	Deep semi-supervised framework	CNN, GIST, SIFT	100-d, 399-d and 1000-d word freq vectors	Wikipedia, pascal VOC, NUS-WIDE	Semi-supervised	MAP	[87]
47	Correspondence autoencoder	Pyramid Histogram of Words (PHOW), MPEG-7 descriptors and Gist	BoW	Wikipedia, Pascal, NUS-WIDE	–	MAP	[86]

(*continued on next page*)

datasets that contain data of various modalities to test and validate the proposed algorithms. The algorithms have been tested on extremely small datasets like the Wikipedia dataset which consists of 2866 documents only. After surveying, it has been found that there is a lack of large multimodal datasets and especially in the medical field [161].

Table 18 (*continued*).

SR.	TECHNIQUE	IMAGE REP.	TEXT REP.	DATA	TYPE	METRIC	REF.
48	Multitask learning approach with 3 modules: Correlation Network, Cross-modal Autoencoder, Latent Semantic Hashing	4096-dimensional vector extracted by the fc10 layer of VGGNet	1386/2000-dimensional bag-of-word vectors	MIRFLICKR-25K, MS COCO	–	MAP	[155]
49	Deep Adversarial Metric Learning (DAML)	SIFT, VGG	LDA, BoW	Wikipedia, Pascal, NUS-WIDE	Supervised	MAP	[97]
50	Deep Pairwise Ranking model with multi-label information for Cross-Modal retrieval (DPRCM)	CNN, GIST, SIFT	100-d, 399-d and 1000-d word freq vectors	Wikipedia, Pascal, NUS-WIDE	Supervised	MAP	[156]
51	Deep Belief Network	LBP	–	NUS-WIDE, Wikipedia	–	MAP, percentage, MRR	[84]
52	Log-Bilinear Model	–	–	IAPR TC-12, attribute discovery, SBU-Captioned photos	–	Bleu, perplexity and retrieval evaluation	[88]

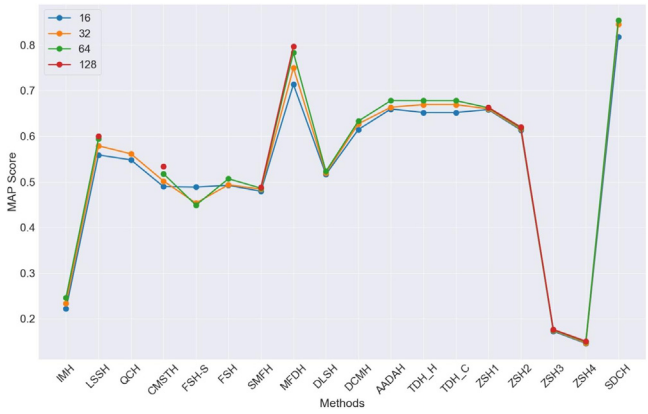

Fig. 25. Average MAP score chart of different hashing methods on NUS-WIDE data.

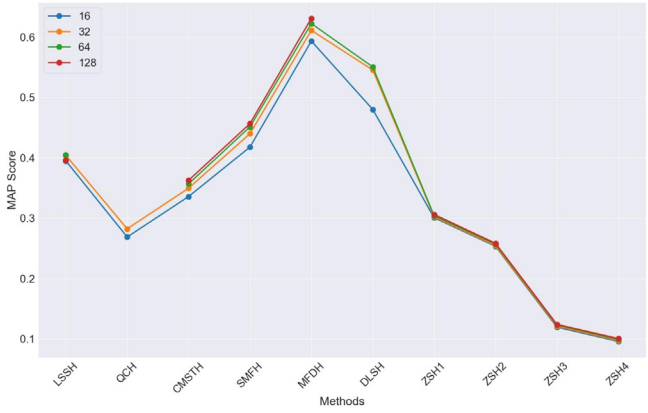

Fig. 26. Average MAP score chart of different hashing methods on Wikipedia data.

4. *Confusion in choosing data feature extraction method*: Most of the approaches used by authors consists of independent feature extraction from each modality to be used in cross-modal system construction as an initial step. If the initial step is inappropriate then it will affect the whole cross-modal technique. As an example, the performance of a machine learning model extremely depends upon the feature representation used for building the model. This happens because various feature representations hide more or less diverse descriptive factors of variations behind the data [162]. So, it is necessary to choose an appropriate feature extraction method for each modality under consideration depending upon the type of data, application, and cross-modal method.
5. *Lack of scalable algorithms*: A colossal amount of multi-modal data is being generated and spread on the internet nowadays due to the availability of fast networks, mobile devices, and huge storage devices. So, productive cross-modal methods are needed which can be applied in a distributed environment as well [23]. Moreover, further research is required for designing efficient cross-modal algorithms which can be applied on huge multi-modal datasets [163].
6. *Need of novel and diverse fields' datasets*: It can be identified from Section 4 and Table 7 that most of the datasets are comprised from social media websites, so their content is very similar to each other. There is an immense need for diversity in the data content. Moreover, the datasets which have been utilized by most of the researchers and are very popular such as NUS-WIDE, Pascal VOC 2007, and Wikipedia, have become too old now. Novel and diverse multi-modal datasets are required to be introduced.
7. *Requirement of semi-supervised cross-modal techniques*: Supervised techniques perform better than unsupervised because of the utilization of semantic label information [116]. However, most of the generated multi-media data is either

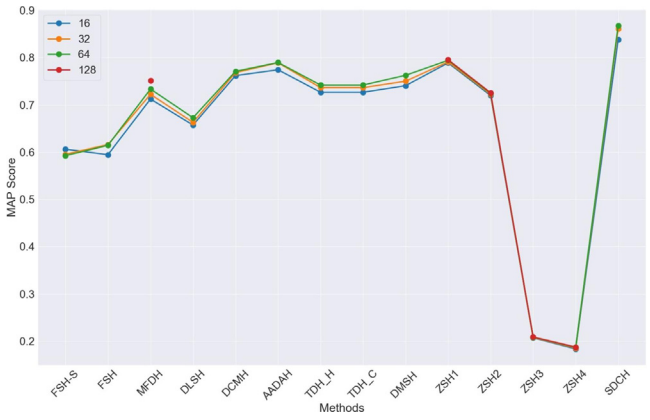

Fig. 27. Average MAP score chart of different hashing methods on MIRFlickr data.

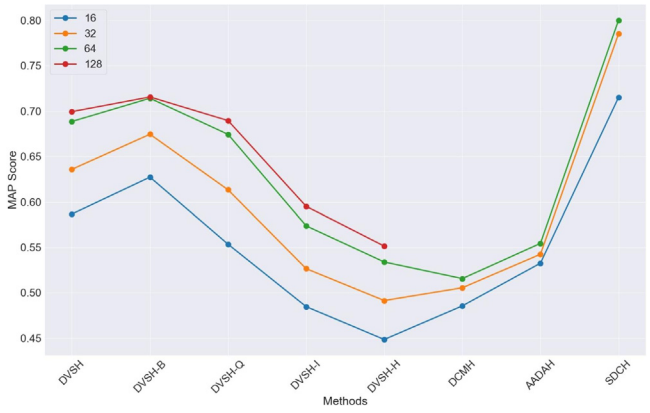

Fig. 28. Average MAP score chart of different hashing methods on IAPR TC-12 data.

unlabeled or has noisy annotations. Semi-supervised methods are getting highly popular now and are the future of cross-modal retrieval as they are the combination of both supervised and unsupervised and also provide promising results [66].

9. Conclusion

From the review on cross-modal information retrieval, it has been found that cross-modal retrieval techniques are better than classic uni-modal systems in retrieving the multi-modal data and adding values to complement meaningful information. The article summarizes the prominent works done by various researchers in the field of image–text cross-modal retrieval. Primary information has been presented with the help of tables, figures, and graphs to make it more understandable. A taxonomy of cross-modal retrieval techniques has been demonstrated. Information regarding famous image–text multi-modal datasets has been presented. Comparison among various cross-modal techniques when applied on a particular dataset is shown. Miscellaneous applications in the field of cross-modal retrieval are mentioned and general architecture is shown. Challenges and open issues have also been discussed to help the research community in further research. Although significant work has been proposed in this field, still we are far away from achieving an ideal position and accuracy in the process. This approach has still not been accepted and applied worldwide in most of the real-life applications. Moreover, ample work is required to be done in this field to introduce novel better algorithms or to enhance the retrieval efficiency of the classic algorithms. It is expected that this article will be useful for the readers and researchers to better understand the present situation and state-of-the-art cross-modal retrieval methods and motivate researches in the field.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] K. Wang, Q. Yin, W. Wang, S. Wu, L. Wang, A comprehensive survey on cross-modal retrieval, 2016, arXiv preprint arXiv:1607.06215.
[2] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.
[3] M. Ayyavaraiah, B. Venkateswarlu, Cross media feature retrieval and optimization: A contemporary review of research scope, challenges and objectives, in: International Conference on Computational Vision and Bio Inspired Computing, Springer, 2019, pp. 1125–1136.
[4] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2372–2385.
[5] M. Ayyavaraiah, B. Venkateswarlu, Joint graph regularization based semantic analysis for cross-media retrieval: a systematic review, Int. J. Eng. Technol. 7 (2.7) (2018) 257–261.
[6] Y.-x. Peng, W.-w. Zhu, Y. Zhao, C.-s. Xu, Q.-m. Huang, H.-q. Lu, Q.-h. Zheng, T.-j. Huang, W. Gao, Cross-media analysis and reasoning: advances and directions, Front. Inf. Technol. Electron. Eng. 18 (1) (2017) 44–57.
[7] M. Priyanka, B. Devi, S. Riyazoddin, M.J. Reddy, Analysis of cross-media web information fusion for text and image association-a survey paper, Global J. Comput. Sci. Technol. (2013).
[8] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, Citeseer, 2007.
[9] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering–a systematic literature review, Inf. Softw. Technol. 51 (1) (2009) 7–15.
[10] B.E. Stein, T.R. Stanford, B.A. Rowland, Development of multisensory integration from the perspective of the individual neuron, Nat. Rev. Neurosci. 15 (8) (2014) 520–535.
[11] R.L. Miller, B.A. Rowland, Multisensory integration: How the brain combines information across the senses, Comput. Model. Brain Behav. (2017) 215–228.
[12] R.K. Srihari, Use of captions and other collateral text in understanding photographs, in: Integration of Natural Language and Vision Processing, Springer, 1995, pp. 245–266.
[13] B.E. Stein, M. Meredith, The Merging of the Senses, The MIT Press, 1993.
[14] B.E. Stein, M.A. Meredith, W.S. Huneycutt, L. McDade, Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli, J. Cogn. Neurosci. 1 (1) (1989) 12–24.
[15] M. Otoom, Beyond von Neumann: Brain-computer structural metaphor, in: 2016 Third International Conference on Electrical, Electronics, Computer Engineering and their Applications, EECEA, IEEE, 2016, pp. 46–51.
[16] B.P. Yuhas, M.H. Goldstein, T.J. Sejnowski, Integration of acoustic and visual speech signals using neural networks, IEEE Commun. Mag. 27 (11) (1989) 65–71.
[17] C. Saraceno, R. Leonardi, Indexing audiovisual databases through joint audio and video processing, Int. J. Imaging Syst. Technol. 9 (5) (1998) 320–331.
[18] D. Roy, Integration of speech and vision using mutual information, in: 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), Vol. 4, IEEE, 2000, pp. 2369–2372.
[19] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (5588) (1976) 746–748.
[20] T. Westerveld, D. Hiemstra, F. De Jong, Extracting bimodal representations for language-based image retrieval, in: Multimedia'99, Springer, 2000, pp. 33–42.
[21] T. Westerveld, Image retrieval: Content versus context, in: RIAO, Citeseer, 2000, pp. 276–284.
[22] C. Xiong, D. Zhang, T. Liu, X. Du, Voice-face cross-modal matching and retrieval: A benchmark, 2019, arXiv preprint arXiv:1911.09338.

[23] A.C. Duarte, Cross-modal neural sign language translation, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 1650–1654.
[24] S. Marioooryad, C. Busso, Exploring cross-modality affective reactions for audiovisual emotion recognition, IEEE Trans. Affect. Comput. 4 (2) (2013) 183–196.
[25] M. Jing, B.W. Scotney, S.A. Coleman, M.T. McGinnity, X. Zhang, S. Kelly, K. Ahmad, A. Schlaf, S. Gründer-Fahrer, G. Heyer, Integration of text and image analysis for flood event image recognition, in: 2016 27th Irish Signals and Systems Conference, ISSC, IEEE, 2016, pp. 1–6.
[26] M.M. Rahman, D. You, M.S. Simpson, S.K. Antani, D. Demner-Fushman, G.R. Thoma, Interactive cross and multimodal biomedical image retrieval based on automatic region-of-interest (ROI) identification and classification, Int. J. Multimed. Inf. Retrieval 3 (3) (2014) 131–146.
[27] Z. Liu, H. Liu, W. Huang, B. Wang, F. Sun, Audiovisual cross-modal material surface retrieval, Neural Comput. Appl. (2019) 1–9.
[28] D. Cao, Z. Yu, H. Zhang, J. Fang, L. Nie, Q. Tian, Video-based cross-modal recipe retrieval, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 1685–1693.
[29] M. Lazaridis, A. Axenopoulos, D. Rafailidis, P. Daras, Multimedia search and retrieval using multimodal annotation propagation and indexing techniques, Signal Process., Image Commun. 28 (4) (2013) 351–367.
[30] D. Xia, L. Miao, A. Fan, A cross-modal multimedia retrieval method using depth correlation mining in big data environment, Multimedia Tools Appl. (2019) 1–16.
[31] X. Zhai, Y. Peng, J. Xiao, Heterogeneous metric learning with joint graph regularization for cross-media retrieval, in: Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.
[32] B. Elizalde, S. Zarar, B. Raj, Cross modal audio search and retrieval with joint embeddings based on text and audio, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 4095–4099.
[33] Y. Yu, S. Tang, F. Raposo, L. Chen, Deep cross-modal correlation learning for audio and lyrics in music retrieval, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 15 (1) (2019) 20.
[34] D. Zeng, Y. Yu, K. Oyama, Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval, 2019, arXiv preprint arXiv:1908.03737.
[35] P. Tripathi, P.P. Watwani, S. Thakur, A. Shaw, S. Sengupta, Discover cross-modal human behavior analysis, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology, ICECA, IEEE, 2018, pp. 1818–1824.
[36] J. Imura, T. Fujisawa, T. Harada, Y. Kuniyoshi, Efficient multi-modal retrieval in conceptual space, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, 2011, pp. 1085–1088.
[37] P. Goyal, S. Sahu, S. Ghosh, C. Lee, Cross-modal learning for multi-modal video categorization, 2020, arXiv preprint arXiv:2003.03501.
[38] J.C. Pereira, N. Vasconcelos, Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems, Comput. Vis. Image Underst. 124 (2014) 123–135.
[39] T. Gou, L. Liu, Q. Liu, Z. Deng, A new approach to cross-modal retrieval, in: Journal of Physics: Conference Series, vol. 1288, no. 1, IOP Publishing, 2019, 012044.
[40] N. Srivastava, R. Salakhutdinov, Learning representations for multimodal data with deep belief nets, in: International Conference on Machine Learning Workshop, Vol. 79, 2012.
[41] Y. Verma, C. Jawahar, A support vector approach for cross-modal search of images and texts, Comput. Vis. Image Underst. 154 (2017) 48–63.
[42] N. Gao, S.-J. Huang, Y. Yan, S. Chen, Cross modal similarity learning with active queries, Pattern Recognit. 75 (2018) 214–222.
[43] A. Habibian, T. Mensink, C.G. Snoek, Discovering semantic vocabularies for cross-media retrieval, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM, 2015, pp. 131–138.
[44] N. Van Nguyen, M. Coustaty, J.-M. Ogier, Multi-modal and cross-modal for lecture videos retrieval, in: 2014 22nd International Conference on Pattern Recognition, IEEE, 2014, pp. 2667–2672.
[45] T. Nakano, A. Kimura, H. Kameoka, S. Miyabe, S. Sagayama, N. Ono, K. Kashino, T. Nishimoto, Automatic video annotation via hierarchical topic trajectory model considering cross-modal correlations, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2011, pp. 2380–2383.
[46] B. Jiang, X. Huang, C. Yang, J. Yuan, Cross-modal video moment retrieval with spatial and language-temporal attention, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, ACM, 2019, pp. 217–225.
[47] X. Xu, L. He, A. Shimada, R.-i. Taniguchi, H. Lu, Learning unified binary codes for cross-modal retrieval via latent semantic hashing, Neurocomputing 213 (2016) 191–203.
[48] K. Ahmad, Slandail: A security system for language and image analysis-project no: 607691, 2017, Available at SSRN 3060047.
[49] A. Hanbury, A survey of methods for image annotation, J. Vis. Lang. Comput. 19 (5) (2008) 617–627.
[50] B. Rafkind, M. Lee, S.-F. Chang, H. Yu, Exploring text and image features to classify images in bioscience literature, in: Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology, Association for Computational Linguistics, 2006, pp. 73–80.
[51] G. Wang, D. Hoiem, D. Forsyth, Building text features for object image classification, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1367–1374.
[52] G. Wang, H. Ji, D. Kong, N. Zhang, Modality-dependent cross-modal retrieval based on graph regularization, Mob. Inf. Syst. 2020 (2020).
[53] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in Statistics, Springer, 1992, pp. 162–190.
[54] C. Guo, D. Wu, Canonical correlation analysis (CCA) based multi-view learning: An overview, 2019, arXiv preprint arXiv:1907.01693.
[55] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Comput. 16 (12) (2004) 2639–2664.
[56] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 251–260.
[57] J.C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G.R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2013) 521–535.
[58] Y. Verma, C. Jawahar, Im2Text and Text2Im: Associating images and texts for cross-modal retrieval, in: BMVC, Vol. 1, Citeseer, 2014, p. 2.
[59] M. Katsurai, T. Ogawa, M. Haseyama, A cross-modal approach for extracting semantic relationships between concepts using tagged images, IEEE Trans. Multimed. 16 (4) (2014) 1059–1074.
[60] J. Shao, Z. Zhao, F. Su, T. Yue, Towards improving canonical correlation analysis for cross-modal retrieval, in: Proceedings of the on Thematic Workshops of ACM Multimedia 2017, 2017, pp. 332–339.
[61] W. Xiong, S. Wang, C. Zhang, Q. Huang, Wiki-cmr: A web cross modality dataset for studying and evaluation of cross modality retrieval models, in: 2013 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2013, pp. 1–6.
[62] V. Ranjan, N. Rasiwasia, C. Jawahar, Multi-label cross-modal retrieval, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4094–4102.
[63] S.J. Hwang, K. Grauman, Accounting for the relative importance of objects in image retrieval, in: BMVC, Vol.1, No. 2, 2010, p. 5.
[64] S. Hwang, K. Grauman, Learning the relative importance of objects from tagged images for retrieval and cross-modal search, Int. J. Computer Vis. 100 (2) (2012) 134–153.
[65] K. Wang, R. He, L. Wang, W. Wang, T. Tan, Joint feature selection and subspace learning for cross-modal retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 38 (10) (2015) 2010–2023.
[66] G. Xu, X. Li, Z. Zhang, Semantic consistency cross-modal retrieval with semi-supervised graph regularization, IEEE Access 8 (2020) 14278–14288.
[67] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Generalized semi-supervised and structured subspace learning for cross-modal retrieval, IEEE Trans. Multimed. 20 (1) (2017) 128–141.
[68] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, S. Yan, Modality-dependent cross-media retrieval, ACM Trans. Intell. Syst. Technol. (TIST) 7 (4) (2016) 1–13.
[69] C. Deng, X. Tang, J. Yan, W. Liu, X. Gao, Discriminative dictionary learning with common label alignment for cross-modal retrieval, IEEE Trans. Multimed. 18 (2) (2015) 208–218.
[70] S. Wang, F. Zhuang, S. Jiang, Q. Huang, Q. Tian, Cluster-sensitive structured correlation analysis for web cross-modal retrieval, Neurocomputing 168 (2015) 747–760.
[71] L. Zhang, B. Ma, G. Li, Q. Huang, Q. Tian, Cross-modal retrieval using multiordered discriminative structured subspace learning, IEEE Trans. Multimed. 19 (6) (2016) 1220–1233.
[72] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 154–162.
[73] G. Cao, A. Iosifidis, K. Chen, M. Gabbouj, Generalized multi-view embedding for visual recognition and cross-modal retrieval, IEEE Trans. Cybern. 48 (9) (2017) 2542–2555.
[74] Y. Wu, S. Wang, G. Song, Q. Huang, Augmented adversarial training for cross-modal retrieval, IEEE Trans. Multimed. (2020).
[75] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, 2003, pp. 119–126.
[76] Y. Xia, Y. Wu, J. Feng, Cross-media retrieval using probabilistic model of automatic image annotation, Int. J. Signal Process. Image Process. Pattern Recognit. 8 (4) (2015) 145–154.

[77] Z. Li, J. Liu, C. Xu, H. Lu, Mlrank: Multi-correlation learning to rank for image annotation, Pattern Recognit. 46 (10) (2013) 2700–2710.
[78] Q. Xu, M. Li, M. Yu, Learning to rank with relational graph and point-wise constraint for cross-modal retrieval, Soft Comput. 23 (19) (2019) 9413–9427.
[79] Y. Wu, S. Wang, Q. Huang, Online fast adaptive low-rank similarity learning for cross-modal retrieval, IEEE Trans. Multimed. (2019).
[80] J. Yu, Y. Cong, Z. Qin, T. Wan, Cross-modal topic correlations for multi-media retrieval, in: Proceedings of the 21st International Conference on Pattern Recognition, ICPR2012, IEEE, 2012, pp. 246–249.
[81] Y. Wang, F. Wu, J. Song, X. Li, Y. Zhuang, Multi-modal mutual topic reinforce modeling for cross-media retrieval, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 307–316.
[82] Z. Qin, J. Yu, Y. Cong, T. Wan, Topic correlation model for cross-modal multimedia information retrieval, Pattern Anal. Appl. 19 (4) (2016) 1007–1022.
[83] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[84] B. Jiang, J. Yang, Z. Lv, K. Tian, Q. Meng, Y. Yan, Internet cross-media retrieval based on deep learning, J. Vis. Commun. Image Represent. 48 (2017) 356–366.
[85] P. Hu, L. Zhen, D. Peng, P. Liu, Scalable deep multimodal learning for cross-modal retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 635–644.
[86] F. Feng, X. Wang, R. Li, I. Ahmad, Correspondence autoencoders for cross-modal retrieval, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 12 (1s) (2015) 26.
[87] D. Mandal, P. Rao, S. Biswas, Semi-supervised cross-modal retrieval with label prediction, IEEE Trans. Multimed. (2019).
[88] R. Kiros, R. Salakhutdinov, R. Zemel, Multimodal neural language models, in: International Conference on Machine Learning, 2014, pp. 595–603.
[89] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 7–16.
[90] F. Feng, R. Li, X. Wang, Deep correspondence restricted Boltzmann machine for cross-modal retrieval, Neurocomputing 154 (2015) 50–60.
[91] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, S. Yan, Cross-modal retrieval with CNN visual features: A new baseline, IEEE Trans. Cybern. 47 (2) (2016) 449–460.
[92] Y. He, S. Xiang, C. Kang, J. Wang, C. Pan, Cross-modal retrieval via deep and bidirectional representation learning, IEEE Trans. Multimed. 18 (7) (2016) 1363–1377.
[93] X. Huang, Y. Peng, M. Yuan, Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval, IEEE Trans. Cybern. (2018).
[94] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, M. Cord, Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 35–44.
[95] J. Gu, J. Cai, S.R. Joty, L. Niu, G. Wang, Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7181–7189.
[96] W. Cao, Q. Lin, Z. He, Z. He, Hybrid representation learning for cross-modal retrieval, Neurocomputing 345 (2019) 45–57.
[97] X. Xu, L. He, H. Lu, L. Gao, Y. Ji, Deep adversarial metric learning for cross-modal retrieval, World Wide Web 22 (2) (2019) 657–672.
[98] X. Xu, H. Lu, J. Song, Y. Yang, H.T. Shen, X. Li, Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval, IEEE Trans. Cybern. (2019).
[99] Z. Yang, Z. Lin, P. Kang, J. Lv, Q. Li, W. Liu, Learning shared semantic space with correlation alignment for cross-modal event retrieval, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 16 (1) (2020) 1–22.
[100] J.-H. Su, C.-L. Chou, C.-Y. Lin, V.S. Tseng, Effective semantic annotation by image-to-concept distribution model, IEEE Trans. Multimed. 13 (3) (2011) 530–538.
[101] L. Chi, X. Zhu, Hashing techniques: A survey and taxonomy, ACM Comput. Surv. 50 (1) (2017) 1–36.
[102] H.P. Luhn, A new method of recording and searching information, Amer. Document. 4 (1) (1953) 14–16.
[103] H. Stevens, Hans Peter Luhn And the birth of the hashing algorithm, IEEE Spectr. 55 (2) (2018) 44–49.
[104] W.W. Peterson, Addressing for random-access storage, IBM J. Res. Dev. 1 (2) (1957) 130–146.
[105] R. Morris, Scatter storage techniques, Commun. ACM 11 (1) (1968) 38–44.
[106] L. Xie, L. Zhu, P. Pan, Y. Lu, Cross-modal self-taught hashing for large-scale image retrieval, Signal Process. 124 (2016) 81–92.
[107] W. Cao, W. Feng, Q. Lin, G. Cao, Z. He, A review of hashing methods for multimodal retrieval, IEEE Access 8 (2020) 15377–15391.
[108] X. Zhu, Z. Huang, H.T. Shen, X. Zhao, Linear cross-modal hashing for efficient multimedia search, in: Proceedings of the 21st ACM International Conference on Multimedia, 2013, pp. 143–152.
[109] B. Wu, Q. Yang, W.-S. Zheng, Y. Wang, J. Wang, Quantized correlation hashing for fast cross-modal search, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.
[110] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, IEEE Trans. Image Process. 27 (8) (2018) 3893–3903.
[111] C. Yan, X. Bai, S. Wang, J. Zhou, E.R. Hancock, Cross-modal hashing with semantic deep embedding, Neurocomputing 337 (2019) 58–66.
[112] X. Lu, L. Zhu, Z. Cheng, X. Song, H. Zhang, Efficient discrete latent semantic hashing for scalable cross-modal retrieval, Signal Process. 154 (2019) 217–231.
[113] H.T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, R. Hong, Exploiting subspace relation in semantic labels for cross-modal hashing, IEEE Trans. Knowl. Data Eng. (2020).
[114] Y. Cao, M. Long, J. Wang, Q. Yang, P.S. Yu, Deep visual-semantic hashing for cross-modal retrieval, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1445–1454.
[115] Q.-Y. Jiang, W.-J. Li, Deep cross-modal hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3232–3240.
[116] J. Yu, X.-J. Wu, J. Kittler, Learning discriminative hashing codes for cross-modal retrieval based on multi-view features, Pattern Anal. Appl. (2020) 1–18.
[117] Y. Fang, H. Zhang, Y. Ren, Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing, Knowl.-Based Syst. 171 (2019) 69–80.
[118] J. Tang, K. Wang, L. Shao, Supervised matrix factorization hashing for cross-modal retrieval, IEEE Trans. Image Process. 25 (7) (2016) 3157–3166.
[119] X. Liu, Z. Li, J. Wang, G. Yu, C. Domeniconi, X. Zhang, Cross-modal Zero-shot Hashing, 2019, arXiv preprint arXiv:1908.07388.
[120] J. Yu, X.-J. Wu, Unsupervised concatenation hashing with sparse constraint for cross-modal retrieval, 2019, arXiv preprint arXiv:1904.00726.
[121] X. Zhang, H. Lai, J. Feng, Attention-aware deep adversarial hashing for cross-modal retrieval, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 591–606.
[122] Y. Gong, S. Lazebnik, A. Gordo, F. Perronnin, Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2012) 2916–2929.
[123] S. Kumar, R. Udupa, Learning hash functions for cross-view similarity search, in: Twenty-Second International Joint Conference on Artificial Intelligence, 2011.
[124] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: Advances in Neural Information Processing Systems, 2009, pp. 1753–1760.
[125] J. Song, Y. Yang, Y. Yang, Z. Huang, H.T. Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013, pp. 785–796.
[126] H. Liu, R. Ji, Y. Wu, F. Huang, B. Zhang, Cross-modality binary code learning via fusion similarity hashing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7380–7388.
[127] X. Shen, F. Shen, Q.-S. Sun, Y.-H. Yuan, H.T. Shen, Robust cross-view hashing for multimedia retrieval, IEEE Signal Process. Lett. 23 (6) (2016) 893–897.
[128] J. Zhou, G. Ding, Y. Guo, Latent semantic sparse hashing for cross-modal similarity search, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, 2014, pp. 415–424.
[129] Z. Ji, W. Yao, W. Wei, H. Song, H. Pi, Deep multi-level semantic hashing for cross-modal retrieval, IEEE Access 7 (2019) 23667–23674.
[130] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, NUS-WIDE: A real-world web image database from national university of Singapore, in: Proc. of ACM Conf. on Image and Video Retrieval, CIVR'09, Santorini, Greece, July 8–10, 2009.
[131] M. Grubinger, P. Clough, H. Müller, T. Deselaers, The iapr tc-12 benchmark: A new evaluation resource for visual information systems, in: International Workshop OntoImage, Vol. 2, 2006.
[132] M. Everingham, L. Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[133] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes challenge 2007 (VOC2007) results, 2007.
[134] M.J. Huiskes, M.S. Lew, The MIR flickr retrieval evaluation, in: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, 2008, pp. 39–43.

[135] M.J. Huiskes, B. Thomee, M.S. Lew, New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative, in: Proceedings of the International Conference on Multimedia Information Retrieval, 2010, pp. 527–536.
[136] J. Krapac, M. Allan, J. Verbeek, F. Juried, Improving web image search results using query-relative classifiers, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 1094–1101.
[137] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, J. Artificial Intelligence Res. 47 (2013) 853–899.
[138] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78.
[139] C. Rashtchian, P. Young, M. Hodosh, J. Hockenmaier, Collecting image annotations using Amazon's Mechanical Turk, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon'S Mechanical Turk, Association for Computational Linguistics, 2010, pp. 139–147.
[140] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[141] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[142] Y. Jia, M. Salzmann, T. Darrell, Learning cross-modality similarity for multinomial data, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2407–2414.
[143] F. Zhong, G. Wang, Z. Chen, F. Xia, G. Min, Cross-modal retrieval for CPSS data, IEEE Access 8 (2020) 16689–16701.
[144] G. Xu, X. Li, L. Shi, Z. Zhang, A. Zhai, Combination subspace graph learning for cross-modal retrieval, Alexandria Eng. J. (2020).
[145] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, Lbmch: Learning bridging mapping for cross-modal hashing, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2015, pp. 999–1002.
[146] G. Ding, Y. Guo, J. Zhou, Y. Gao, Large-scale cross-modality search via collective matrix factorization hashing, IEEE Trans. Image Process. 25 (11) (2016) 5427–5440.
[147] X. Zhang, K. Ahmad, Ontology and terminology of disaster management, in: DIMPLE: DIsaster Management and Principled Large-Scale Information Extraction Workshop Programme, 2014, p. 46.
[148] M. Rogers, K. Ahmad, Corpus Linguistics and Terminology Extraction, John Benjamins, 2001.
[149] Z. Zhongming, L. Linong, Y. Xiaona, Z. Wangqiang, L. Wei, et al., WIKI-CMR: A web cross modality database for studing and evaluation of cross modality retrieval methods, 2013.
[150] C. Kang, S. Xiang, S. Liao, C. Xu, C. Pan, Learning consistent feature representation for cross-modal multimedia retrieval, IEEE Trans. Multimed. 17 (3) (2015) 370–381.
[151] L. Wu, Y. Wang, L. Shao, Cycle-consistent deep generative hashing for cross-modal retrieval, IEEE Trans. Image Process. 28 (4) (2018) 1602–1612.
[152] Y. Peng, X. Huang, J. Qi, Cross-media shared representation by hierarchical learning with multiple deep networks, in: IJCAI, 2016, pp. 3846–3853.
[153] J. Shao, L. Wang, Z. Zhao, A. Cai, et al., Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval, Neurocomputing 214 (2016) 618–628.
[154] V.E. Liong, J. Lu, Y.-P. Tan, J. Zhou, Deep coupled metric learning for cross-modal matching, IEEE Trans. Multimed. 19 (6) (2016) 1234–1244.
[155] J. Luo, Y. Shen, X. Ao, Z. Zhao, M. Yang, Cross-modal image-text retrieval with multitask learning, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2309–2312.
[156] Y. Jian, J. Xiao, Y. Cao, A. Khan, J. Zhu, Deep pairwise ranking with multi-label information for cross-modal retrieval, in: 2019 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2019, pp. 1810–1815.
[157] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, S. Belongie, Learning from noisy large-scale datasets with minimal supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 839–847.
[158] C. Tian, V. De Silva, M. Caine, S. Swanson, Use of machine learning to automate the identification of basketball strategies using whole team player tracking data, Appl. Sci. 10 (1) (2020) 24.
[159] D.J. Armaghani, G.D. Hatzigeorgiou, C. Karamani, A. Skentou, I. Zoumpoulaki, P.G. Asteris, Soft computing-based techniques for concrete beams shear strength, Proced. Struct. Integrity 17 (2019) 924–933.
[160] C. Raghuraman, S. Suresh, S. Shivshankar, R. Chapaneri, Static and dynamic malware analysis using machine learning, in: First International Conference on Sustainable Technologies for Computational Intelligence, Springer, 2020, pp. 793–806.
[161] H. Müller, D. Unay, Retrieval from and understanding of large-scale multi-modal medical datasets: A review, IEEE Trans. Multimed. 19 (9) (2017) 2093–2104.
[162] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[163] Y. Jia, L. Bai, S. Liu, P. Wang, J. Guo, Y. Xie, Semantically-enhanced kernel canonical correlation analysis: a multi-label cross-modal retrieval, Multimedia Tools Appl. 78 (10) (2019) 13169–13188.