A cross-modal system for image annotation and retrieval

A Thesis submitted in fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

Submitted by

Parminder Kaur
(Registration no: 901803021)

Under the supervision of

Dr. Husanbir Singh Pannu

Assistant Professor, CSED, TIET

Dr. Avleen Kaur Malhi

Assistant Professor, University of Warwick, UK

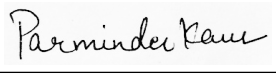
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala-147004, Punjab, India

September 2023

Certificate

I, Parminder Kaur, hereby declare that the work, which is being presented in the thesis, entitled "*A cross-modal system for image annotation and retrieval*", in fulfillment of the requirements for the award of the degree of *Doctor of Philosophy* in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Husanbir Singh Pannu* and *Dr. Avleen Kaur Malhi*.

The matter presented in this thesis has not been submitted for award of any other degree of this or any other University.



Parminder Kaur
Regn. No. 901803021

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Dr. Husanbir Singh Pannu
Assistant Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology

Dr. Avleen Kaur Malhi
Assistant Professor
Warwick Manufacturing Group
University of Warwick, UK

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 or cross-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.

Lately, cross-modal retrieval has attained plenty of attention due to enormous multi-modal data generation every day in the form of audio, video, image, and text. One vital requirement of cross-modal retrieval is to reduce the heterogeneity gap among miscellaneous modalities so that one modality's results can be effectively retrieved from the other. So, a novel unsupervised cross-modal retrieval framework (association of image and text modalities) based on associative learning is proposed in this thesis where two traditional SOMs are trained separately for images and collateral text and then they are integrated together using the Hebbian learning network to facilitate the cross-modal retrieval process. Experimental outcomes on a popular Wikipedia dataset and the primary endoscopy data demonstrate that the presented technique outshines various existing state-of-the-art techniques.

Acknowledgement

First and foremost, I would like to thank the almighty God who gave me strength and courage to overcome all the obstacles and complete this endeavor. The successful completion of any task would be incomplete without acknowledging the people who made it possible. I would like to take this opportunity to express my gratitude to all those who made this journey possible. Words are often too less to express one's deepest regards, but lets give it a go.

I offer my sincerest gratitude to my supervisors, **Dr. Husanbir Singh Pannu** and **Dr. Avleen Kaur Malhi**, who have supported me throughout my Ph.D. work with their patience and knowledge; while providing me the room to work in my own way. They led me to the correct direction at every stage of this research work. Apart from providing me with excellent supervision, strong cooperation and constant encouragement throughout this journey, they also shared their invaluable experiences with me to succeed in life.

I am also grateful to the head of department, **Dr. Shalini Batra**, the former head, **Prof. Maninder Singh**, Ph.D. coordinator, **Dr. Sushma Jain**, Ph.D. co-coordinator, **Dr. Vinay Arora**, and the members of my doctoral committee, **Dr. Rupali Bhardwaj**, **Dr. Shreelekha Pandey** and **Dr. Vinay Kumar** for their constructive suggestions and ensuring the correct pace of my work. I sincerely thank all the faculty, specially, **Dr. Prashant Singh Rana** and **Dr. Harpreet Singh** and the support staff of Computer Science and Engineering Department as well for their constant help whenever required. I am also obliged to the Director, **Prof. Prakash Gopalan**, Dean (RSP), **Prof. Rafat Siddique** and the management of Thapar Institute, who provided me with all the necessary resources and facilities to complete my work.

The chain of my gratitude will definitely be incomplete if I forget to thank my family for their unconditional love, support and encouragement in every phase of my life. The journey of Ph.D. has been a sweet and bitter ride at times but my family stood by me through thick and thin, and gave me courage at the times when I felt really low. My mother's constant motivation showed me the silver lining in the dark clouds.

I would also like to thank my friends and colleagues with whom I have traveled this journey of research. These people have made my research journey all the more memorable and pleasant. As one cannot mention the names of all well-wishers, friends and beloved ones, I would like to pay my regards to one and all who supported me during this journey of knowledge.

Parminder Kaur

List of Publications

1. Parminder Kaur, Husanbir Singh Pannu, Avleen Kaur Malhi, "Comparative analysis on cross-modal information retrieval: A review", ***Computer Science Review***, Volume 39, February 2021, 100336. *[IF: 12.9]*

2. Parminder Kaur, Avleen Kaur Malhi, Husanbir Singh Pannu, "Hybrid SOM based cross-modal retrieval exploiting Hebbian learning", ***Knowledge-Based Systems***, Volume 239, March 2022, 108014. *[IF: 8.8]*

3. Parminder Kaur, Avleen Kaur Malhi, Husanbir Singh Pannu, "Sentiment analysis of linguistic cues to assist medical image classification", ***Multimedia Tools and Applications***, September 2023, 1-20. *[IF: 3.6]*

4. Parminder Kaur, Avleen Kaur Malhi, Husanbir Singh Pannu, "Annotate and retrieve in-vivo images using hybrid self-organizing map", ***The Visual Computer***. *[Accepted, IF: 3.5]*

5. Parminder Kaur, Avleen Kaur Malhi, Husanbir Singh Pannu, "Image clustering using Zernike moments and self-organizing maps for gastrointestinal tract", ***Complex & Intelligent Systems***. *[Revision submitted, IF: 5.8]*

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AADAH	Attention-Aware Deep Adversarial Hashing
ACMR	Adversarial Cross-Modal Retrieval
AICDM	Annotation by Image-to-Concept Distribution Model
BoVW	Bag of Visual Words
BoW	Bag of Words
CCA	Canonical Correlation Analysis
CCDCR	Class Center Discriminant analysis for Cross-modal Retrieval
CM	Correlation Matching
CM_l1	CM with l1 distance measure
CM_l2	CM with l2 distance measure
CM_NC	CM with normalized correlation measure
CM_NC_c	CM with centered normalized correlation measure
CMOLRS	Cross-Modal Online Low-Rank Similarity
CMSTH	Cross-Modal Self-Taught Hashing
CMTC	Cross-Modal Topic Correlation
CNN	Convolutional Neural Network
Corr-AE	Correspondence Autoencoder
CR-CDSL	Cross-modal Retrieval with Collective Deep Semantic Learning
CSGL	Combination Subspace Graph Learning
CVH	Cross-View Hashing
DAML	Deep Adversarial Metric Learning
$DAML\_D$	Deep Adversarial Metric Learning with deep features
$DAML\_S$	Deep Adversarial Metric Learning with shallow features
DCMH	Deep cross-modal hashing
DDL	Discriminative Dictionary Learning
DLA-CMR	Adversarial Cross-Modal Retrieval based on Dictionary Learning
DLSH	Discrete Latent Semantic Hashing
DMSH	Deep Multi-level Semantic Hashing
DVSH	Deep Visual Semantic Hashing
DVSH-B	DVSH variant without binarization
DVSH-H	DVSH variant without using the hashing networks
DVSH-I	DVSH variant by replacing the cosine max-margin loss
DVSH-Q	DVSH variant without bitwise max-margin loss
FSH	Fusion Similarity Hashing
FSH-S	FSH with a simple fusion graph construction
GSS-SL	Generalized Semisupervised Structured Subspace learning

HOG	Histogram of Oriented Gradients
HRL	Hybrid Representation Learning
HRL_C	Hybrid Representation Learning with CNN features
HRL_H	Hybrid Representation Learning with handcrafted features
IMH	Inter Media Hashing
JFSSL	Joint Feature Selection and Subspace Learning
KCCA	Kernel Canonical Correlation Analysis
KDM	Kernel Dependence Maximization
LBP	Local Binary Pattern
LCMH	Linear Cross-Modal Hashing
LDA	Latent Dirichlet Allocation
LSSH	Latent Semantic Sparse Hashing
M3R	Multi-Modal Mutual topic Reinforcement modelling
MAP	Mean Average Precision
MDCR	Modality Dependent Cross-media Retrieval
MDSSL	Multiordered Discriminative Structured Subspace Learning
MFDH	Multi-view Feature Discrete Hashing
MHTN	Modal-adversarial Hybrid Transfer Network
MJSL	Multi-class Joint Subspace Learning
MLRank	Multi-correlation Learning to Rank
MRR	Mean Reciprocal Rank
MSFH	Multi-modal graph regularized Smooth matrix Factorization Hashing
PR curve	Precision Recall curve
QCH	Quantized Correlation Hashing
RCH	Robust Cross-view Hashing
S3CA	Shared Semantic Space with Correlation Alignment
SCCMR	Semantic Consistency Cross-Modal Retrieval
SCM	Semantic Correlation Matching
SCM_KL	SCM with Kullback-Leibler divergence measure
SCM_l1	SCM with l1 distance measure
SCM_l2	SCM with l2 distance measure
SCM_NC	SCM with normalized correlation measure
SCM_NC_c	SCM with centered normalized correlation measure
SDCH	Semantic Deep Cross-modal Hashing
SGRCR	Supervised Graph Regularization based Cross-media Retrieval
SIFT	Scale Invariant Feature Transformation
SM	Semantic Matching
SM_KL	SM with Kullback-Leibler divergence measure
SM_l1	SM with l1 distance measure

SM_l2	SM with l2 distance measure
SM_NC	SM with normalized correlation measure
SM_NC_c	SM with centered normalized correlation measure
SMDCR	Semi-supervised Modality-Dependent Cross-media Retrieval
SMFH	Supervised Matrix Factorization Hashing
SRLCH	Subspace Relation Learning for Cross-modal Hashing technique
SVM	Support Vector Machine
TDH	Triplet based Deep Hashing
TDH_C	TDH with CNN-F features
TDH_H	TDH with handcrafted features
TQSL	Task-dependent and Query-dependent Subspace Learning
UCH	Unsupervised Concatenation Hashing
UCH_LLE	UCH with Locally Linear Embedding
UCH_LPP	UCH with Locality Preserving Projection
ZSH	Zero-Shot Hashing
ZSH1	ZSH (with complete data)
ZSH2	ZSH (with zero shot data)
ZSH3	ZSH (with semi-supervised zero shot data)
ZSH4	ZSH (with semi-supervised zero shot data and different label spaces)

Chapter 1
Introduction

In real life, data is often represented in miscellaneous forms or comprised of diverse domains. Therefore, the data associated with the same underlying object, event, or content may exist as different modalities and exhibit heterogeneous properties. For instance, when one visits a new location then he records the memory by taking pictures, recording videos, or posting a piece of microblog. Although all these data forms represent the same event, they are considered separate modalities. When we are unable to completely understand the content depicted in an image embedded in the text, then picture captions or referral text often help. For example, after looking at a volleyball image (Figure 1.1), one may not be able to exactly recognize the ball or know about the volleyball game. However, the image can be easily recognized with the help of some collateral text present in the form of figure reference, caption, and related citation. This implies that information from multiple sources or in diverse forms is always helpful for better understanding.

Volleyball is a team sport in which two teams of six players are separated by a net. Each team tries to score points by grounding a ball on the other team's court under organized rules as studied in **[1]. Figure 1** shows the picture of volleyball.

Figure 1: Picture of volleyball

Figure 1.1: An example of a volleyball image and collateral text in the form of the caption, figure reference, and related citation.

1.1 Information retrieval

In simple terms, information retrieval is the process of obtaining the required results from a database through a system based upon the input query. It can broadly be classified into three types: uni-modal, cross-modal, and multi-modal. 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. 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 matched 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. Uni-modal, cross-modal and multi-modal systems are explained using a simple example in Figure 1.2 where + represents both text and images can be retrieved using an image query and vice versa in multi-modal approach. Multimodality provides a vital property known as complementarity, which means each modality brings some kind of added value to the whole system that cannot be obtained from other modalities present in the system [4]. This added value is the diversity provided by each modality which contributes to the overall efficiency of the multi-modal system by enhancing interpretability, uniqueness, and robustness.

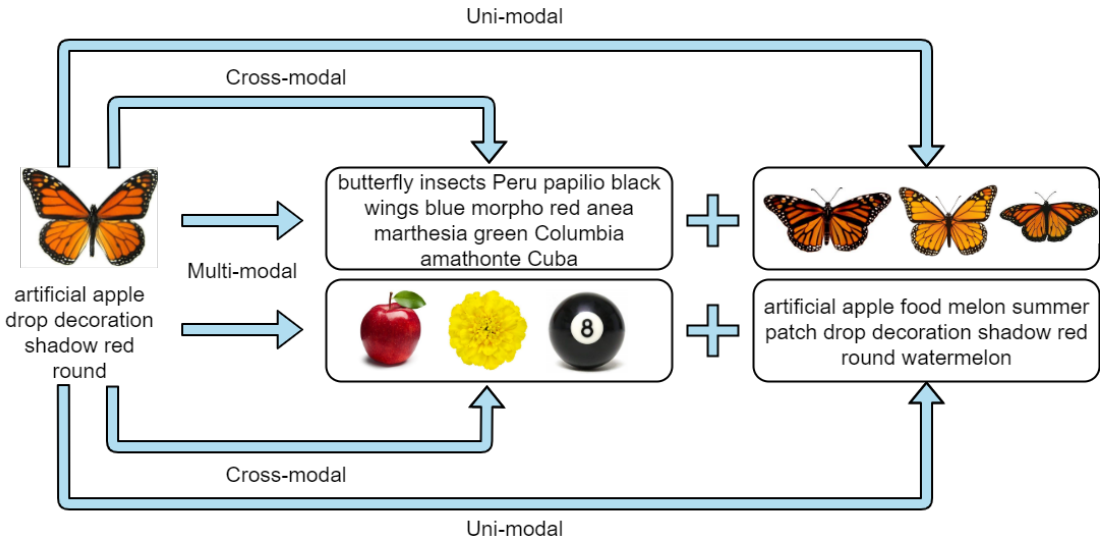


Figure 1.2: An illustration of information retrieval in uni-modal, cross-modal, and multi-modal system.

With the advent of the internet and multiple social media websites, a massive amount of data in the form of diverse modalities such as images, videos, audio, and text keeps increasing daily. With an increase in this type of heterogeneous data, searching and extracting appropriate (and inter-modal) information becomes necessary but difficult [5]. Classic uni-modal information retrieval techniques are of the least use these days as they are unable to handle the enormous amount of multi-media data. There is a requirement for creating novel systems capable of handling the multi-modal data, indexing and organizing it in a way such that they can be utilized for efficient information retrieval comprising various modalities. The systems should also be capable of fusing information from different modalities

and inferring valuable results. So, there is a need for a cross-modal or multi-modal information fusion and retrieval system which can work effectively in the case of cross-domains and specifically, retrieval of one modality using another modality. Therefore, the fundamental idea of cross-modal is to integrate numerous modes of information to derive better results than just one channel.

1.1.1 Image annotation and retrieval

In simple words, image annotation is a process of explaining an image with appropriate linguistic cues. It can also be defined as retrieving the text or set of keywords from a group/database of texts that best describes an image query. 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 underconstrain 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.

1.2 Background

The inception of the terms *cross-modal* and *multi-modal* is in neurology and are inspired from multi-sensory integration inside brain [6, 7]. Humans familiarize themselves with the environment using different sensory modalities such as vision, taste, and hearing where each modality provides a distinctive impression of the surroundings [8]. Each sensory mode works separately to connect with the environment and acquire information. This information obtained through the modalities is then integrated inside the brain for providing atmosphere-related awareness and to reach an inference or take a suitable action [9]. For instance, if someone is unable to get the meaning of his interlocutor then he will subsequently start noticing his body and facial expressions to accurately understand his message. Likewise,

information in one modality is inadequate for proper understanding of an event or an object. Thus, a concept of information fusion encompassing image and text is studied in this research which is inspired by the working of the human brain. Figure 1.3 illustrates the process of information fusion inside brain. Firstly, the brain acquires information in different forms from the surroundings through various senses or sensory modalities such as sight, touch, hearing, taste, and smell. Afterward, it associates (fuses) all the forms of data along with prior experiences or knowledge to make a decision or perform a further action as per the input data. Similarly, a multi-modal system fuses information from different modalities which can be utilized for multiple applications such image/video captioning [10, 11], cross-modal information retrieval [12], emotion and sentiment classification [13], and visual question answering [14].

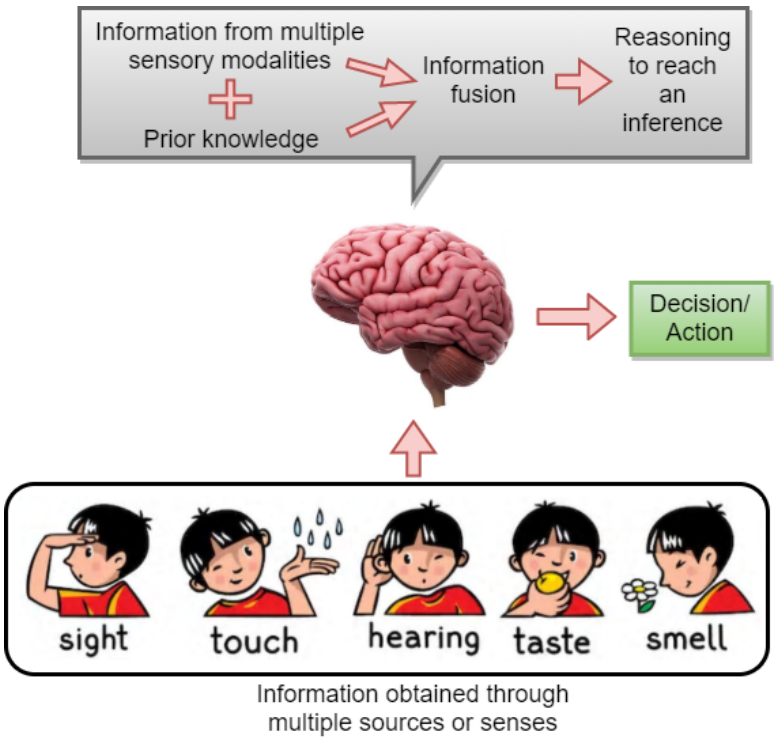


Figure 1.3: Simple illustration of information fusion inside brain

There are three types of multi-modal or cross-modal fusion techniques: early fusion, late fusion, and intermediate fusion which are defined as follows [4, 15]:

1. *Early fusion* or *Data-level fusion* is a classic way of fusing or integrating diverse modalities before carrying out the analysis. It applies to raw or pre-processed data obtained from various modalities. Modalities should be represented by feature vectors before fusion. It is difficult to synchronize the heterogeneous data sources where one data source could be discrete and the other continuous. Thus, the major challenge in this type of fusion is

the conversion of diverse data sources into a single feature vector. Early fusion process is represented in Figure 1.4. An ample amount of data will be removed from the modalities to create a common ground before fusion, which is a major drawback of this approach.

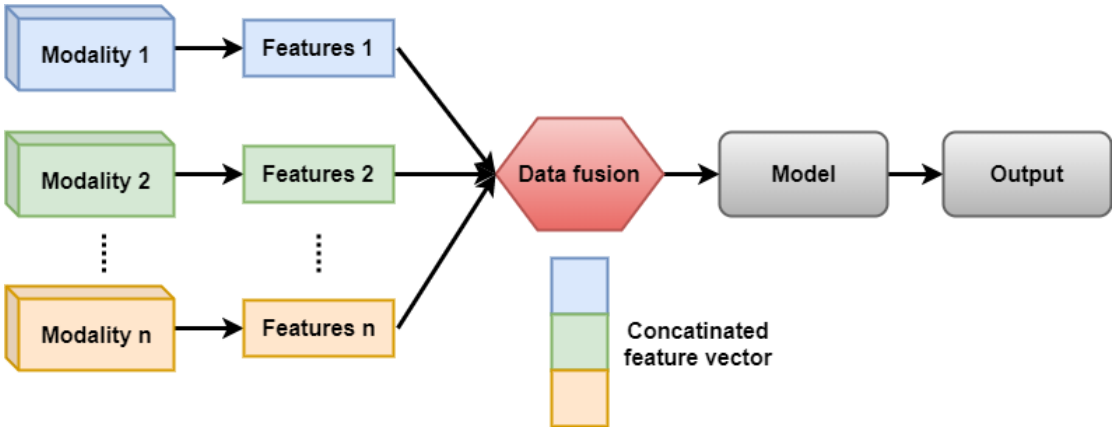


Figure 1.4: Multi-modal early fusion technique

2. In *Late fusion* or *Decision-level fusion*, data fusion happens at the last stage, just before retrieving the output. Separate models are trained for individual feature vectors. Output from the trained models is fused at a decision-making stage as depicted in Figure 1.5. This technique became famous with the popularity of the ensemble classifiers [16]. This approach is straightforward compared to early fusion, especially when the modalities are remarkably different in terms of data dimensionality and sampling rate.

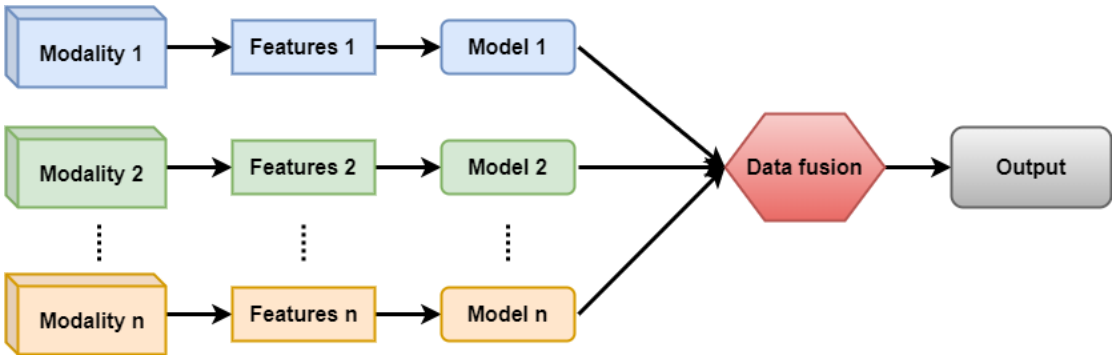


Figure 1.5: Multi-modal late fusion technique

The errors from multiple models corresponding to different modalities are handled independently, so this approach performs better than the early fusion approach. However, there is no definite proof of this as per [17]. Still, many researchers utilize the late fusion approach for analyzing their respective multi-modal data problems.

3. *Intermediate fusion* allows data fusion at different phases or depths of model training which makes this technique very flexible. It is built based on deep neural networks, and their performance has also been enhanced. This fusion changes the input data into a higher level of representation as it passes through different layers. Each layer provides a new representation of the input modalities by applying linear and nonlinear functions. In the context of multi-modal deep learning, immediate fusion is the association of miscellaneous modalities' representations into a single layer such that the model learns the joint representation of the modalities. The data fusion of modalities can be performed simultaneously or gradually with one or multiple modalities at a time. Numerous modality feature vectors can be associated in a single layer, however, it may cause model overfitting, or the network may not be able to learn the proper relationships among the modalities. Figure 1.6 depicts an instance of the intermediate fusion technique where modalities' features are combined in diverse network layers.

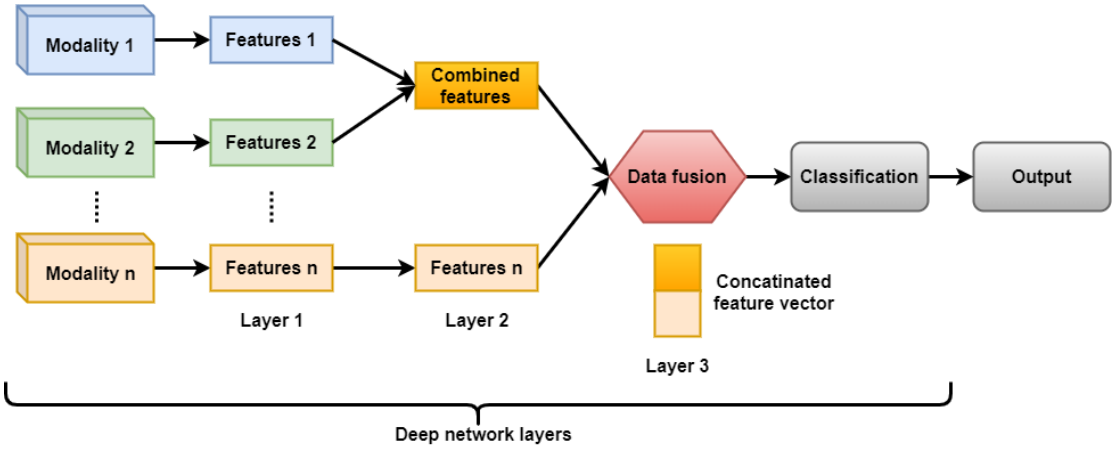


Figure 1.6: Multi-modal intermediate fusion technique

Most of the initial works which are inspired by cross-modal human behavior are related to the integration of acoustic and visual modalities. In [18], authors have associated the acoustic with visual speech signals for automatic speech recognition. The experimentation has been performed on the visual speech signals of a male using neural networks, and the initial set of results is limited to the static images of vowels only. The results demonstrated that visual speech signals provide supplementary information in addition to only acoustic signals and help in effective vowel recognition along with the presence of acoustic noise.

The article [19] focuses on representing and organizing the audio-visual information for browsing and retrieval tasks. It has been presented that the audio-visual sequence can be easily organized into semantically meaningful segments

using cross-modal analysis of simple audio and visual information. These segments represent different scenes that are coherent from a semantic viewpoint. The results in this article have demonstrated that audio classification plays a vital role in forming connections among consecutive shots, which allows for obtaining a scene-level description. When the non-consecutive shots correlate, a higher level of abstraction can also be achieved, known as video idioms. A generic audio model has been proposed that categorizes the audio signal into four element types: silence, music, speech, and numerous other sounds. A system has been developed in [20] to learn words from visual input accompanied by spoken input. The objective is to split the continuous speech automatically at word boundaries without a lexicon and to create visual classes corresponding to the spoken words. Mutual information combines the visual and acoustic distance metrics to extract an audio-visual lexicon from raw input. The experimentation has been performed on a corpus comprising infant-directed speech and images.

Numerous works proposed in the 80s and 90s are influenced by the McGurk effect [21]. 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. The McGurk effect can be experienced using a phoneme's production video accompanied by a sound recording of another phoneme being spoken. For instance, the syllables *ba-ba* spoken over the lip movements of *ga-ga* creates a perception of *da-da*. The authors initially believed that the reason for this effect was the common phonetic and visual characteristics of "b" and "g". Two kinds of illusion has been observed regarding incompatible audio-visual stimuli: fusions (auditory *ba* and visual *ga* creates *da*) and integrations (auditory *ga* and visual *ba* creates *bga*). The brain is guessing at its best regarding the incoming information. The information coming from the ears and eyes is contradictory. Still, the eyes or the visual information have more impact on the brain, and hence the fusion and integration responses have been generated.

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 [22, 23]. 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.

1.2.1 Cross-modal retrieval architecture

Figure 1.7 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 raw data contains noise that affects the overall building and accuracy of the system. So the data is pre-processed to remove that noise and to make it appropriate for further processing on it. In the second step each modality is represented separately using miscellaneous feature extraction 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.

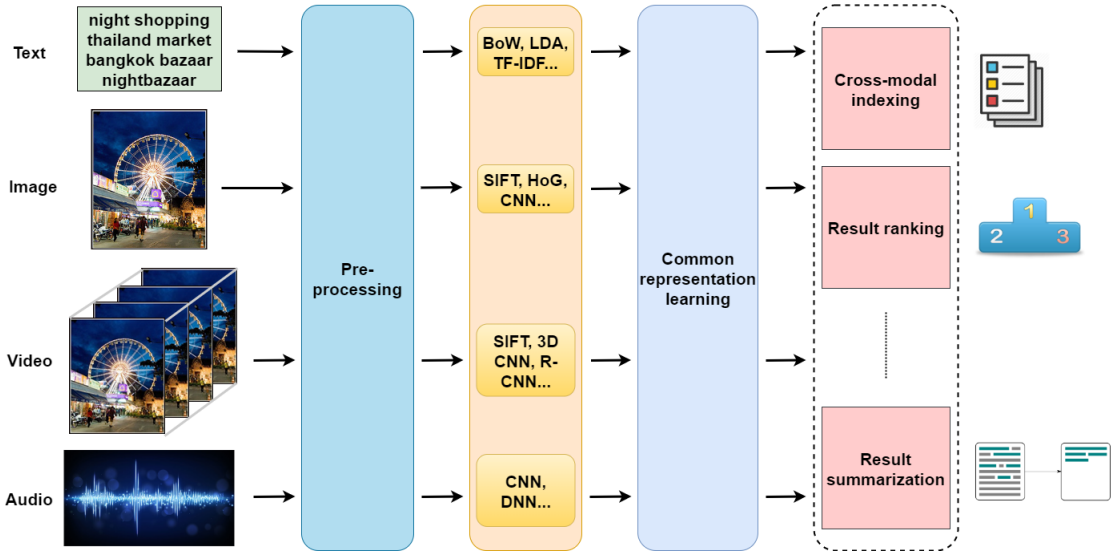


Figure 1.7: General framework of cross-modal retrieval process

Figure 1.8 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 unlabelled 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

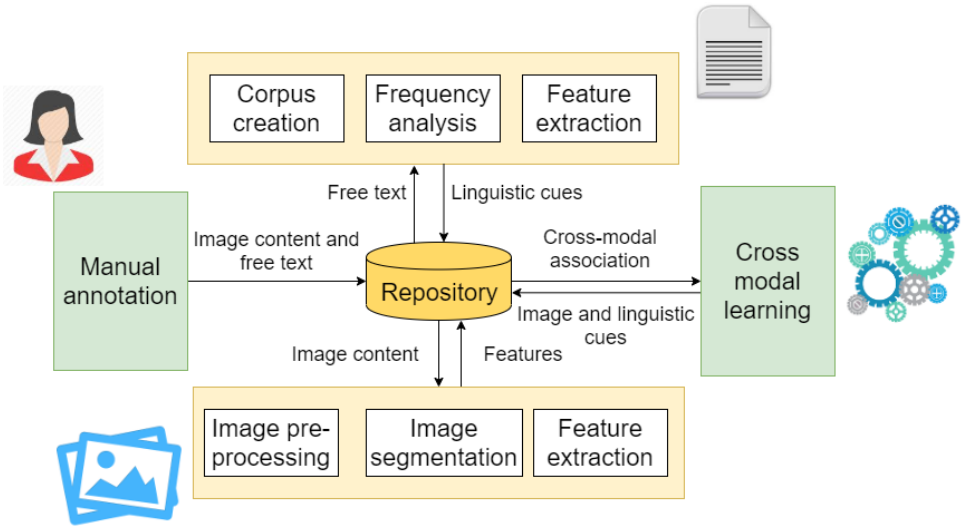


Figure 1.8: General topological architecture of image-text cross-modal system

and text analysis modules separately for analysis and then the extracted features are put back in the repository. The cross-modal learning module 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.

1.3 Motivation

Querying images using visual features is usually not viable when the images involve technical information or deliberate reasoning, for example, medical image diagnosis, forensic investigation, or space science images. There is often a semantic gap between visual features and objects portrayed in it. Therefore collateral text must be associated with the visual learning system for completeness of the information. Current cross-modal systems such as Google for text/image, YouTube video recommender systems, Amazon shopping advice, are also not perfect for their retrieval accuracy. It is often observed that with the increase in the number of search keywords, the retrieved text results become spurious. In the case of image retrieval using a textual query with extra keywords, the performance becomes even worse. It often suffers from the repetition of previous results and irrelevant information.

Figure 1.9: (a) Burning Notre Dame Cathedral (b) 9/11 attack at World Trade Centre New York City. In both the images, texture features for the smoke are identical and therefore additional linguistic information must be associated with images to classify them in separate categories i.e. accidental destruction by fire versus terrorist attacks with planes hijacked.

YouTube is a popular video-sharing platform and even it is not 100% accurate in recommending videos and association of related multimedia. As an example, a gigantic fire accident happened at the well-known Notre Dame Cathedral in Paris

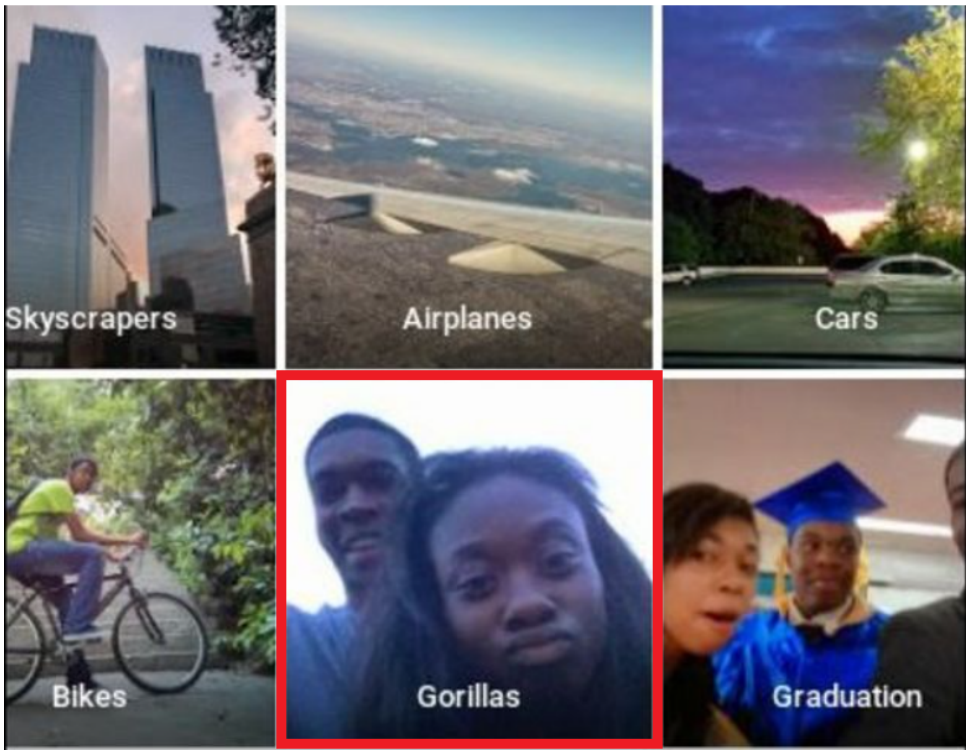


Figure 1.10: Google's Photos app did a racist blunder

on 15^{th} *April*2019 reported by *abcNEWS* [24]. YouTube mistakenly related this incident with the 9/11 terrorist attack at World Trade Center in New York City as shown in Figure 1.9. Many videos of the Cathedral burning incident showed the text about 9/11 from Encyclopedia Britannica. Later on, YouTube apologized for this mistake. This mistake might have happened due to similarity in smoke texture in both the images. Another real life example includes the labeling of images of black people as *gorillas* by Google's *Photos* app as shown in Figure 1.10. Google apologized later for this racist blunder, reported by BBC News on 1^{st} *July* 2015 [25]. As a solution to this accidental algorithmic racism issue, Google banned words such as *gorillas*, *chimpanzee*, and *monkey*, reported by The Guardian on 12^{th} *January* 2018 . Similarly, there are a number of real life examples like these. Hence information representing the same entity in other forms such as linguistic cues is required to compensate the missing information in a single modality. Thus, a novel cross-modal retrieval system has been proposed in this research that is inspired by the working of human brain and capable of retrieving information in one modality using another modality.

1.4 Challenges

The major challenges observed in the process of cross-modal retrieval are represented in Figure 1.11 and defined as follows:

1. *Massive multimedia data sets*: Sophisticated content retrieval has become a challenge in ever-growing multimedia data volume over the internet [26]. 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 [27]. 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 [28].

4. *Need of efficient feature extraction*: Choosing optimal feature extraction method for underlying multi-modal data is still an open question [29], [30]. 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 [31].

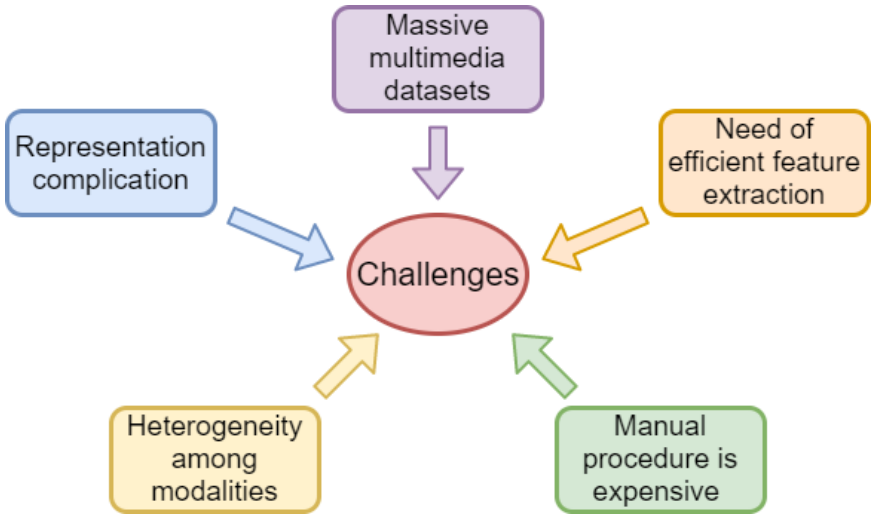


Figure 1.11: Challenges in cross-modal retrieval

1.5 Applications

Multi-modal information fusion and cross-modal retrieval has miscellaneous application areas. However, below are the few recent applications considering only image and text modalities and these are also depicted in Figure 1.12.

1. *Image-recipe retrieval and generation*: Food holds a vital place in humans' life as it is associated with our health, culture, and feelings. In [32], authors have introduced a huge recipe dataset incorporating over one million cooking recipes and thirteen million related food images. A neural network has been trained to learn the joint embedding of food images and corresponding recipes that facilitates the task of image-recipe retrieval and also demonstrates remarkable results. Authors have proposed an online recipe generation and evaluation system using the same recipe dataset in [33]. There are two modes of text generation in this system: (a) generation of the ingredient list from a recipe name and its cooking instructions; and (b) generation of cooking steps from a recipe title and the list of ingredients.

2. *Histopathological multi-modal retrieval*: Cancer is a universal health issue and one of the major reasons for deaths around the globe. The technological advancement has enabled the composition of an enormous collection of digital histopathology slides along with clinical information and other metadata.

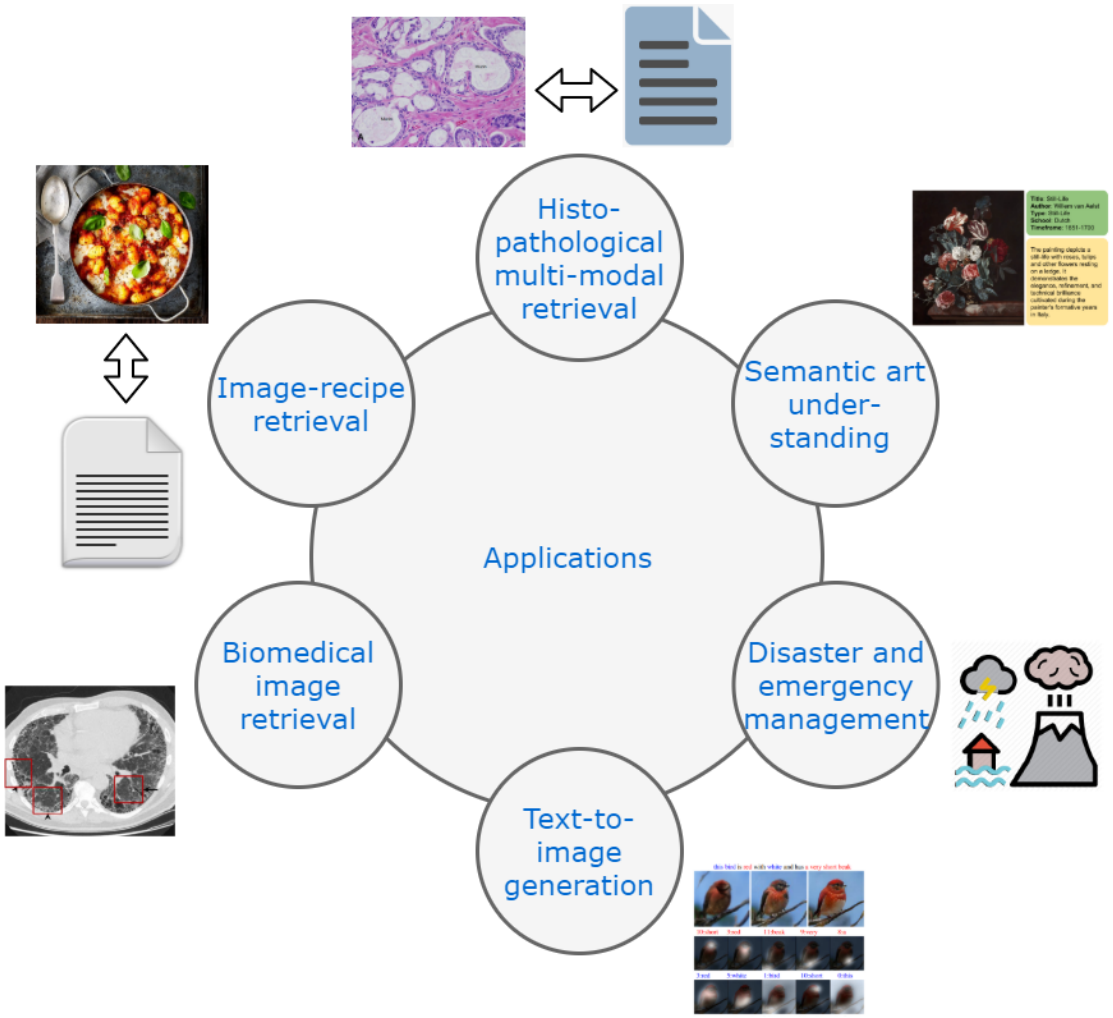


Figure 1.12: Few multi-modal applications incorporating image and text modalities

This collection of data can act as a valuable source for prognosis, diagnosis of cancer, and theragnosis purposes. Authors in [34] have utilized this data compilation and proposed a novel multi-modal retrieval approach that is based on a supervised multi-modal kernel semantic embedding model. The introduced system is capable of searching the relevant prostate cancer cases in a multi-modal data collection including both histology slides (images) and pathologist's reports (text).

3. *Semantic art understanding*: Understanding an artistic demonstration comprises various intricate procedures such as detecting author influences or recognizing the components in the scene. A new multi-modal dataset dubbed *SemArt* has been introduced in [35] that contains fine-art painting images along with various features and a textual artistic comment. A cross-modal system has been designed and trained for retrieving suitable painting images by using an artistic textual query and vice versa. Two context-aware em-

bedding techniques have been proposed in [36] for automatic art analyzing which includes type categorization, author identification, and cross-modal retrieval by utilizing the *SemArt* dataset.

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 [37] to extract prominent phrases related to floods. A 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 Facebook groups and pages for the flood and the aid (in German).

5. *Text to image generation*: Automatic image generation from natural language explanation is a fundamental issue in various applications such as computer-aided design and art generation. An Attentional Generative Adversarial Network (AttnGAN) technique has been proposed for fine-grained and attention-driven generation of an image from text [38]. AttnGAN can produce fine details at different sub-sections of an image by attentively analyzing the textual description to identify the significant words.

6. *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 [39] 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.

1.6 Thesis contribution

The prominent contributions of this thesis can be broadly classified into below three levels:

- At a *philosophical/ psychological level* we imitate the working of neurons inside the brain. We are introducing new ways of intelligently indexing and querying image collections by training neural computing systems with images that have collateral texts. More precisely, given the two representations of the same event or entity, this thesis focuses on introducing a way to extract the missing information in one representation (modality) from the complementary modality. Moreover, the cross-modal interaction between both modalities enables the trained neural system to retrieve the information different from the query modality, such as retrieving related texts using an image query and vice versa.

- At a *theoretical level* this research focuses on integrating unsupervised neural networks to generate multi-net systems with a goal of scrutinizing whether the integration of independently trained neural networks can learn to effectively categorize and extract information that is imparted in diverse modalities compared to single-net or uni-modal system. For this implementation, two self-organizing maps (SOM) are trained separately on image features and collateral text features and then associated with the Hebbian network to create the final cross-modal information retrieval system. We evaluate the robustness of the proposed cross-modal approach against the uni-modal system (comprising a traditional SOM) and the stat-of-the-art cross-modal information retrieval techniques incorporating image and text modalities.

- At an *application level* the major contribution of the thesis is in the area of automatic image annotation and text illustration. The proposed approach learns the association between diverse modalities, so it has the capability of effective classification and multi-modal information retrieval (comprising retrieval results from the same or different modality than the query). The significant contribution of this research is of unsupervised image-text cross-modal system. In reality, getting labeled/annotated data is quite tricky, so the proposed framework will work effectively in that case as it is unsupervised and thus does not require any data labeling.

1.7 Thesis organization

The thesis has been organized in various chapters as follows:

- **Chapter 1: Introduction**
  This chapter provides background and an introduction to cross-modal and multi-modal information retrieval and the motivation behind this research. A general cross-modal information retrieval architecture along with topological architecture has been shown. It also presents the various challenges faced by the researchers while implementing a cross-modal retrieval framework and the applications in the field of image-text cross-modal retrieval. Moreover, it also describes the thesis contribution at *philosophical, theoretical*, and *application* level. In the end, it presents the thesis organization.

- **Chapter 2: Literature review**
  This chapter summarizes the work done by miscellaneous researchers in the field of image-text cross-modal information retrieval. Moreover, it provides a comparative analysis of the existing cross-modal techniques in a tabular form. The popular benchmark datasets and the evaluation metrics for evaluating the cross-modal retrieval system have also been covered. Moreover, the chapter discusses the research gaps identified after reviewing the cross-modal and multi-modal literature. In the end, the objectives of the proposed study are presented.

- **Chapter 3: Hybrid self-organizing map approach**
  This chapter presents the implementation of the proposed cross-modal retrieval framework: hybrid SOM exploiting Hebbian learning ($HSOM\_HEBB$). Firstly, the raw image and text data are pre-processed separately, and appropriate feature extractors are utilized for extracting features from each modality. Secondly, two individual self-organizing maps (SOMs) are trained with these extracted features. Ultimately, these two trained SOMs are integrated using an associative Hebbian network.

- **Chapter 4: Image clustering using WT, ZM, and SOM**
  This chapter comprises a detailed explanation of the endoscopy image data analysis and the clustering procedure. Wavelet Transform (WT) has been utilized for image noise removal before feature extraction. Then the image features are extracted using the Zernike moments (ZM), and the clustering of the in-vivo gastrointestinal images is performed using a self-organizing map (SOM).

- **Chapter 5: Cross-modal retrieval using Oja rule**
  This chapter proposes the $HSOM\_OJA$ method which is an extension of the technique explained in chapter 3 with application in endoscopy data. Here, deep visual features are extracted using a pre-trained VGG16 convolution neural network, and TFIDF is used for text feature extraction. Two separate SOMs are trained using these features and then integrated using improved Hebb links (Hebb rule) or Oja links (Oja rule).

- **Chapter 6: Experimental analysis**
  This chapter includes the two case studies' experiments, evaluation, and results. The first case study is the application of the proposed HSOM approach exploiting Hebbian learning ($HSOM\_HEBB$) on the popular Wikipedia dataset and comparison with state-of-the-art methods. It also incorporates the comparison of traditional SOM with hybrid SOM based on information retrieval. The second case study includes the application of the proposed HSOM approach exploiting the Oja rule ($HSOM\_OJA$) on the gastrointestinal images and the collateral text for cross-modal information retrieval. The chapter compares the proposed $HSOM\_HEBB$ and $HSOM\_OJA$ techniques on the endoscopy dataset.

- **Chapter 7: Conclusion and future scope**
  This chapter finally concludes the thesis work by highlighting the significant contributions made toward the proposed research domain. Moreover, the chapter also provides the limitations of the proposed research and the future directions of the work carried out in the thesis.

Chapter 2
Literature Review

Cross-modal representation and retrieval techniques can be predominantly categorized into two classes: (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. Figure(2.1) presents a taxonomy of cross-modal retrieval methods.

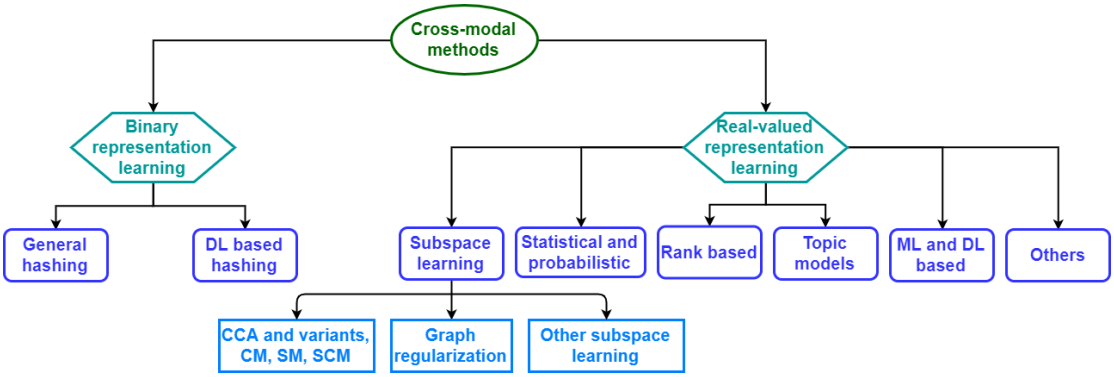


Figure 2.1: Taxonomy of cross-modal retrieval methods

2.1 Real-valued representation learning

This section presents the information regarding various real-valued representation learning methods and their application on diverse datasets.

2.1.1 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 [1]. The modalities can be mapped to common isomorphic

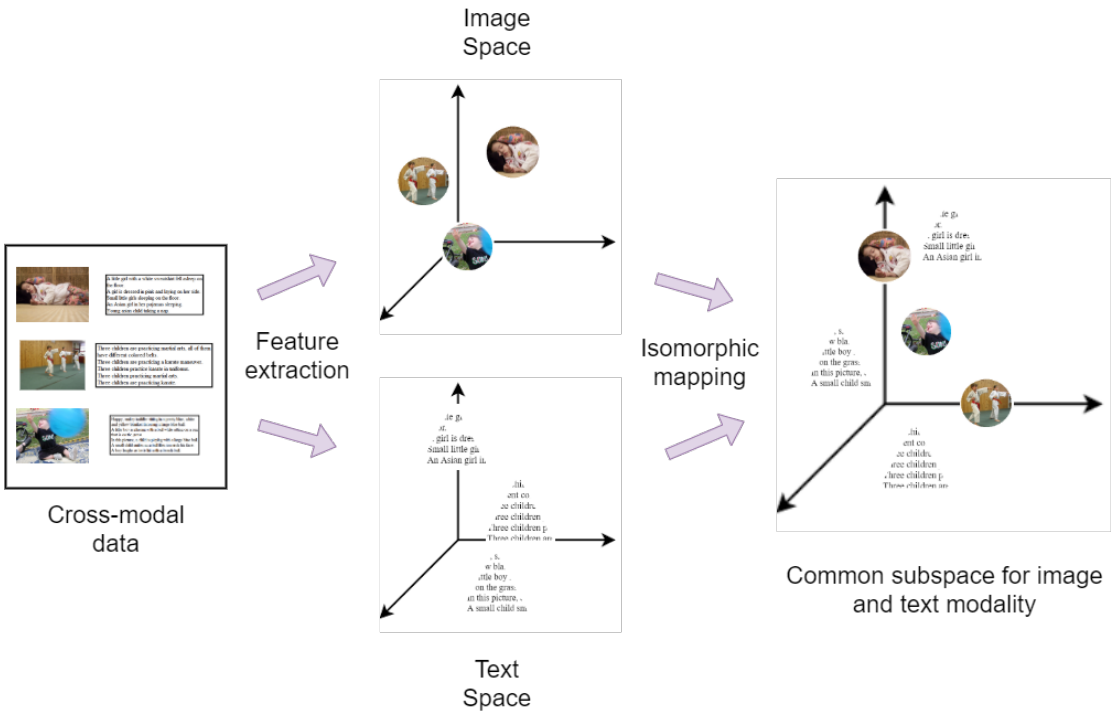


Figure 2.2: General representation of subspace learning process

subspaces from old miscellaneous spaces by learning potential common subspaces (as shown in figure 2.2).

CCA and its variants, CM, SM and SCM

CCA is one of the oldest and the most popular unsupervised technique of subspace learning which was introduced by Hotelling [40] 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 [41]. 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 [42]. 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 $S = ((p_1, q_1), ..., (p_n, q_n))$. S_p represents $(p_1, ..., p_n)$ and $S_q = (q_1, ..., q_n)$, 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: $S_{p,d_p} = (\langle d_p, p_1 \rangle, ..., \langle d_p, p_n \rangle)$ and similarly $S_{q,d_q} = (\langle d_q, q_1 \rangle, ..., \langle d_q, q_n \rangle)$. First step is to choose d_p and d_q for maximizing the correlation between vectors, such that:

$$\rho = \max_{d_p, d_q} Corr(S_p d_p, S_q d_q) = \max_{d_p, d_q} \frac{\langle S_p d_p, S_q d_q \rangle}{\|S_p d_p\| \, \|S_q d_q\|} \tag{2.1}$$

where ρ represents the equation to be maximized. Let E denotes the empirical

expectation of function $f(p, q)$ and given by

$$E = \frac{1}{m} \sum_{i=1}^{m} f(p_i, q_i) \tag{2.2}$$

then ρ can be redefined as

$$\rho = \max_{d_p, d_q} \frac{E[\langle d_p, p \rangle \langle d_q, q \rangle]}{\sqrt{E[\langle d_p, p \rangle^2] E[\langle d_q, q \rangle^2]}} \tag{2.3}$$

$$= \max_{d_p, d_q} \frac{E[d'_p p q' d_q]}{\sqrt{E[d'_p p p' d_p] E[d'_q q q' d_q]}} \tag{2.4}$$

$$= \max_{d_p, d_q} \frac{d'_p E[p q'] d_q}{\sqrt{d'_p E[p p'] d_p d'_q E[q q'] d_q}} \tag{2.5}$$

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

$$Cov(p, q) = E\left[\binom{p}{q} \binom{p}{q}' \right] = \begin{bmatrix} C_{pp} & C_{pq} \\ C_{qp} & C_{qq} \end{bmatrix} = C \tag{2.6}$$

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. 2.6 represents the covariance matrix in zero-mean case only. Now ρ can be redefined as:

$$\rho = \max_{d_p, d_q} \frac{d'_p C_{pq} d_q}{\sqrt{d'_p C_{pp} d_p d'_q C_{qq} d_q}} \tag{2.7}$$

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

Rasiwasia et al. have proposed the use of CCA, Semantic Matching (SM), and Semantic Correlation Matching (SCM) in cross-modal document retrieval task [43]. Two hypotheses have been investigated in this: (a) Benefit to explicitly modelling the correlation between two elements, and (b) This modelling 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 [44] which associates cross-modal re-

trieval 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.

Verma et al. presented a cross-modal retrieval framework 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 [45]. 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 [46] is an extension of [45]. 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. Katsurai et al. has proposed a cross-modal technique for extracting semantic relationship between classes using annotated images [47]. 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 [41]: (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 [48]. To handle the drawbacks of classic CCA, several variants of this method are introduced such as Generalized CCA (GCCA), Kernal 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. Xiong et al. have introduced a new dataset containing images, text (paragraph), and hyperlinks [49]. 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 [48] 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. Ranjan et al. have introduced an extension of the CCA approach, named multi-label CCA (ml-CCA) [50]. 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 [51]. 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 [52] 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 [31]. 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. Wang et al. are dealing with both the issues [53]. 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.

An optimal solution for cross-modal retrieval is provided by [54] which combines the label prediction and optimization of projection matrices into an integrated framework. The method dubbed semantic consistency cross-modal retrieval with semi-supervised graph regularization (SCCMR) makes use of the semantic information present in un-annotated data. It utilizes graph embedding for considering the nearest neighbors in a potential subspace of text and images and text and images having the same semantics.

A combination subspace graph learning (CSGL) supervised cross-modal retrieval approach is proposed in [55]. An objective function is incorporated with graph regularization for original data structure-preserving in projective space. A collaborative learning strategy is utilized for eluding suboptimal solutions while optimization. CSGL technique makes use of semantic information and the original modality distribution. A supervised graph regularization based cross-media retrieval (SGRCR) approach is proposed in [56] that includes learning of two couples of projections as per separate retrieval exercises. Heterogeneous and isomorphic adjacent graphs are built for preserving cross-media data correlations. A class center discriminant analysis for cross-modal retrieval (CCDCR) method has been introduced in [57] which is based on graph regularization. For enhancing the discriminant capability of the method, an inter-modality distance of class center samples is minimized and intra-modality distance is maximized.

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, Wang et al. have proposed a cross-modal retrieval framework which is based on graph regularization and modality dependence, fully utilizing the correlation between modalities [1]. 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 Figure 2.3. The objective function for I2T and T2I tasks are defined in Equations 2.8 and 2.9 respectively.

$$\begin{aligned} F(U_1, V_1) = \lambda \left\| U_1^T X - V_1^T Y \right\|_F^2 + (1-\lambda) \left\| U_1^T X - S \right\|_F^2 + \\ \alpha tr(U_1 X^T L_1 X U_1^T - S^T L_1 S) + \\ \beta_1 \|U_1\|_F^2 + \beta_2 \|V_1\|_F^2 \end{aligned} \tag{2.8}$$

$$\begin{aligned} F(U_2, V_2) = \lambda \left\| U_2^T X - V_2^T Y \right\|_F^2 + (1-\lambda) \left\| V_2^T Y - S \right\|_F^2 + \\ \alpha tr(V_2 Y^T L_2 Y V_2^T - S^T L_2 S) + \\ \beta_1 \|U_2\|_F^2 + \beta_2 \|V_2\|_F^2 \end{aligned} \tag{2.9}$$

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 [54] 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.

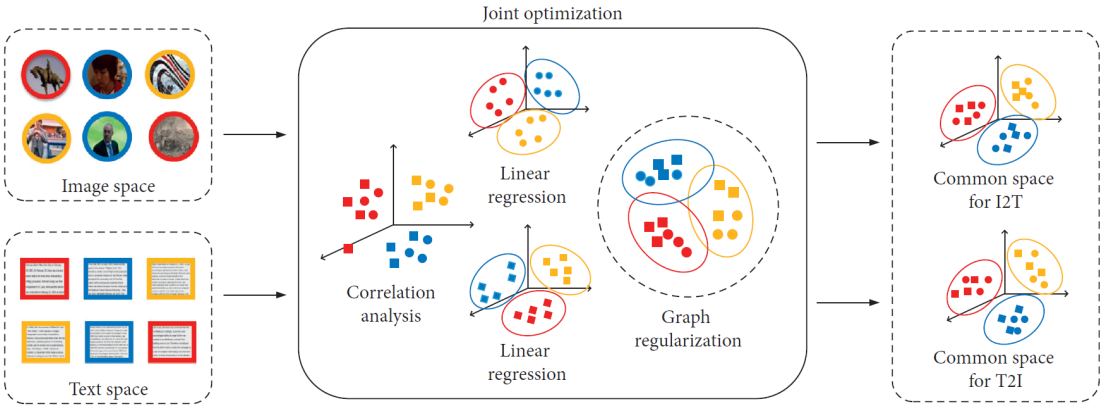


Figure 2.3: Process of cross-modal retrieval framework followed in [1]

Inspired by the fact that unlabelled data can be composed easily and aid to exploit the correlation between modalities, Zhang et al. have proposed a novel framework generalized semi-supervised structured subspace learning (GSS-SL) for cross-modal retrieval [58]. 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 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 [59] 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 [60] 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, Wang et al. have proposed a cluster-sensitive cross-modal correlation learning framework [61]. 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 [62]. 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 [63] 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 [64]. 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 [58] 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 [65]. An augmented version of the conditional generative adversarial network is utilized for reserving the semantic meaning in the modality transfer process. An adversarial cross-modal retrieval based on dictionary learning (DLA-CMR) framework is introduced in [66]. Adversarial learning extracts the arithmetic features of every modality and dictionary learning aids as feature re-constructor for reconstructing discerning features.

2.1.2 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 [39] 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 [67] 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 [67] is presented in [68] 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.

2.1.3 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. Li et al. proposed a new Multi-correlation Learning to Rank (MLRank) approach for image annotation which ranks the tags for images as per their relevance after considering semantic importance and visual similarity [69]. Two cases are defined: (a) image-bias consistency; and (b) tag-bias consistency that is developed into an optimization problem for rank learning.

Xu et al. optimized a ranking model 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 [70]. 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 [71] 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.

2.1.4 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 [72] which jointly models the text and image modalities. A statistical correlation model is examined which is conditioned on category information. Wang et al. 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 [73].

A topic correlation model (TCM) is presented in [74] 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.

2.1.5 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 (DL) mimics the way the human brain works for both feature extraction and classification as discussed in [75]. 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 2.6. Srivastava et al. have proposed a novel technique of multi-modal Deep Belief Network for finding out the missing data in text or image modality [76]. 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 [77] 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-Marquart 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 [78]. They have utilized Local Binary Pattern (LBP) as an image descriptor and Deep Belief Network (DBN) as a deep learning algorithm.

Hu et al. introduced a new Scalable Deep Multi-modal Learning (SDML) data

retrieval method [79]. 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 [80] 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 [81] 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. Kiros et al. 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 [82]. 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 [83] 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 [84] 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 [85]. A deep semantic matching (DSM) technique is also introduced for handling cross-modal retrieval w.r.t. samples labeled with one or multiple labels. He et al. have proposed a deep and bidirectional representation learning model (DBRLM) where images and text are represented by two separate convolutional based networks [86].

A novel modal-adversarial hybrid transfer network has been proposed in [87]. 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 [88], named, AdaMine (ADAptive MINing Embedding) for learning the common representation of recipe items incorporating recipe images and their recipe in textual form. Gu et al. 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 [89]. A deep neural network based approach known as hybrid representation learning (HRL) is proposed for learning common representation for each modality [90].

A new deep adversarial metric learning (DAML) technique is introduced for cross-modal retrieval which maps annotated data pairs of diverse modalities nonlinearly into a shared latent feature subspace [91]. 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, Xu et al. have presented a ternary adversarial network with self-supervision (TANSS) model [92]. 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. Yang et al. proposed a shared semantic space with correlation alignment (S3CA) for cross-modal data representation [93]. It aligns the non-linear correlations of cross-modal data distribution in deep neural networks made for diversified data.

Zhang et al. have presented a novel cross-modal retrieval with collective deep semantic learning (CR-CDSL) approach which makes use of two complementing deep neural networks and deep restricted Boltzmann machines are utilized for weight initialization in the neural networks [94].

Associative learning based approaches

This section presents a short summary of multi-modal retrieval methods that are primarily focused on the philosophies of cognitively logical ways of constructing representations consistent with the inherent re-constructive and associative essence of human memory. So, these methods are inspired by the multi-modal sensory integration inside the human brain. The work presented in this paper falls under this category where an associative Hebbian network has been utilized to integrate two SOMs (image SOM and test SOM) so that the accuracy of cross-modal retrieval can be enhanced by combining two diverse information sources representing the same content.

Shriwas et al. have introduced the use of auto-associative Hopfield network for the cross-modal retrieval process [95]. Experiments have been carried out on image-caption data and the system is tested for various kinds of queries like caption only, image only, and image+caption. The network's retrieval robustness for content-addressable multi-modal pattern retrieval has been assured. Multi-modal associative learning has been introduced in [96] with the use of a modified hypernetwork model or layered hypernetwork. The model comprises two layers incorporating two modality-specific hypernetworks and one modality combining hypernetwork. Korean magazine articles have been utilized for conducting experiments. Hypernetwork association model has also been used in [97] where a vertex denotes a visual patch or a textual word ad hyperedge indicates a higher-order multi-model linkage. Sequential Bayesian sampling has also been exploited in the multi-modal hypernetwork based retrieval of images using text.

A multiSOM approach has been proposed in [98]. The working of the tradi-

tional SOM has been extended for handling different modalities and developing bidirectional associations between them. Heterogeneous data from diverse modalities are associated using the available semantic data as a medium. The multi-modal data considered for experimentation includes images, the human voice of Chinese characters, and their meanings as semantic information. It is ensured that the model learns the bidirectional associative relationship. Wermter et al. have proposed an associative self-organizing framework to integrate multi-modal inputs of vision, language and motor programs for producing complex robot behaviors [99].

Collell et al. have presented a novel approach for constructing multi-modal representations by learning a language-to-vision mapping and its result has been used to create multi-modal embeddings [100]. Authors guarantee that the proposed method acts in an associative and re-constructive way close to human memory. Motivated by the associative and reconstructive nature of human memory, a new associative multichannel autoencoder (AMA) approach has been proposed in [101]. The issue of learning multi-modal word representations by linking visual, auditory, and textual inputs has been considered.

2.1.6 Other methods

This section includes the summary of those works which cannot be classified under any of the above classes. Su et al. In 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 [102]. 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, Jing et al. have performed image retrieval enhancement in the field of floods and flood aids [37]. 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 [103], the accuracy of semantic features is improved using cross-modal regularization which is based on associated text. A task-

dependent and query-dependent subspace learning (TQSL) method has been proposed in [104]. Firstly, a task and category-specific subspaces are learned together in an integrated cross-modal learning framework using an iterative optimization. A task category projection matrix is made based on the previous step. Afterward, a semantic mapping function between multi-modal documents and corresponding classes is learned via a trained linear classifier. Motivated from Hilvert space theory, Xu et al. have proposed a correlation-based cross-modal subspace learning model using kernel dependence maximization (KDM) [105]. Subspace representation for a modality is learned by increasing the kernel dependence rather than direct maximization of feature correlations across multi-modal data. A multi-class joint subspace learning (MJSL) approach is presented in [106] which distinguishes among diverse concepts and utilizes the shared data related to semantic overlap. Dong et al. have presented a semi-supervised modality-dependent cross-media retrieval (SMDCR) method [107]. SMDCR completely utilizes the global data distribution property and semantic data related to both labeled and unlabeled samples.

2.2 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 [108]. The idea of hashing was first introduced by *H. P. Luhn* in his 1953 article *A new method of recording and searching information* [109]. Entire information regarding the birth of hashing is presented in [110]. 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 [111]. Then the term *hashing* was first used by *Morris* in his article [112] in 1968. Few general definitions in hashing are described below [108]:

- *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. $A = [a_1, a_2, ..., a_n] \in R^D$ (real coordinate space of dimension D) and a hashing function $h(\cdot)$, then $h(A) = [h(a_1), h(a_2), ..., h(a_n)] \in [0, p]$ 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 neighbour (NN)*: It represents one or more data entities in $A = [a_1, a_2, ..., a_n] \in R^D$ which are nearest to the query point a_q.
- *Approximate nearest neighbour (ANN)*: It attempts to find a data point $a_x \in A$ which is an $\varepsilon-$approximate nearest neighbour 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 [113]. 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 [114]. 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 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.

Table 2.1 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 [115]. 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 [2]. The distance metric is the metric utilized in the inter or intra similarity among modalities' calculation.

Table 2.1: 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 [116], QCH [117]
		D	TDH [118], SDCH [119]
	Discrete	T	DLSH [120], SRLCH [2]
		D	DVSH [121], DCMH [122]
	Alternative solution	T	MFDH [123], MSFH [115], SMFH [124]
		D	ZSH [125]
Hash function	Linear	T	UCH [126], LCMH [116], CMSTH [113], MSFH [115]
	Non-linear	T	SRLCH [2]
		D	DVSH [121]
Distance metric	Cosine	T	QCH [117]
		D	DVSH [121]
	Euclidean	T	LCMH [116], MSFH [115]
		D	ZSH [125]
	Hamming	T	DLSH [120], CMSTH [113]
		D	DCMH [122], TDH [118], AADAH [127]

2.2.1 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. Yu et al. 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 [126]. $l_{2,1}$-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 [128] as well. CVH [129] is an extension of classic uni-modal spectral hashing [130] to multi-modal field. In IMH [131], learned binary codes conserve both inter and intra-media consistency. FSH [132] embeds the graph-based fusion similarity to a common hamming space. In RCH [133], common hamming space is learned in diverse modalities' binary codes are created as consistent as possible.

Xie et al. have introduced Cross-Modal Self-Taught Hashing (CMSTH) technique for both cross-modal and uni-modal image retrieval [113]. 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 [116] 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:

$$\begin{aligned} &\min_{B^{(1)},B^{(2)}} \left\|B^{(1)} - B^{(2)}\right\|_F^2; s.t., {B^{(i)}}^T e = 0, \\ &b^{(i)} \in \{-1, 1\}, {B^{(i)}}^T B^{(i)} = I_c, i = 1, 2; \end{aligned} \tag{2.10}$$

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 $\left\|B^{(1)} - B^{(2)}\right\|_F^2$ obtains the maximal consistency (or the minimal difference) on two object representations. Equation 2.10 is extended for more than two modality case and the new general equation obtained is:

$$\begin{aligned} &\min_{B^{(i)},i=1,...,p} \sum_{i=1}^{p} \sum_{i<j}^{p} \left\|B^{(i)} - B^{(j)}\right\|_F^2; \\ &s.t., {B^{(i)}}^T e = 0, b^{(i)} \in \{-1, 1\}, {B^{(i)}}^T B^{(i)} = I_c, i = 1, ..., p, \end{aligned} \tag{2.11}$$

where p represents no. of diverse modalities and rest of the notations are same as

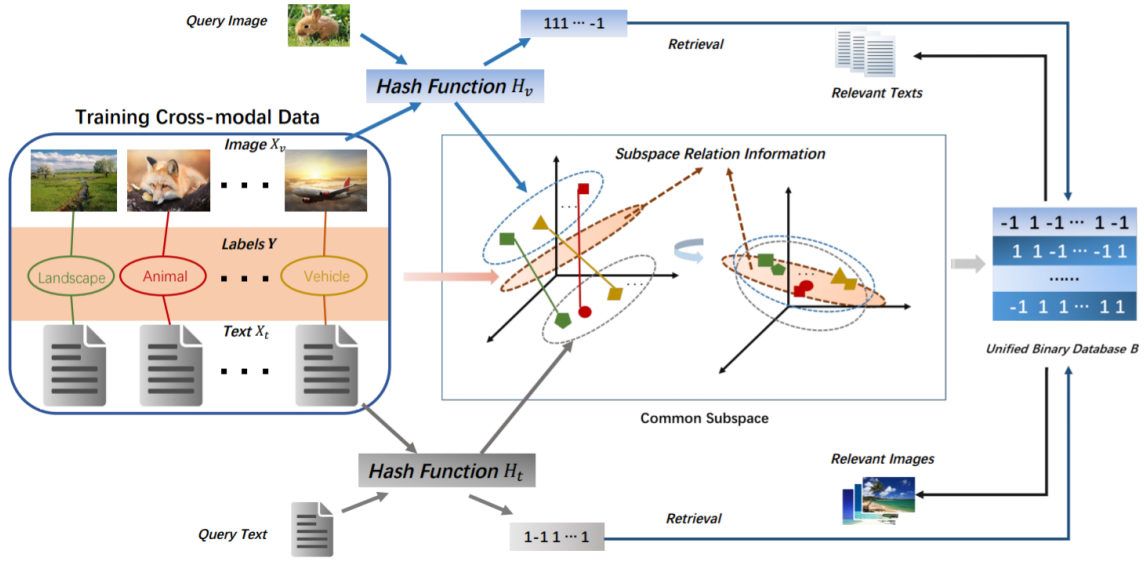


Figure 2.4: Cross-modal hashing approach proposed in [2]

eq. (2.10).

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 [134]. 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.

Wu et al. proposed a quantized correlation hashing (QCH) technique which considers the quantization loss over different modalities and the relation among them simultaneously [117]. 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 Equation 2.12. Suppose two modalities (x_i, y_i) are representing n object, where x_i^T depicts i^{th} row of data matrix $X \in R^{n \times d_x}$ of one modal and y_i^T represents i^{th} 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.

$$\begin{aligned} &min\, O(B_x, B_y, W_x, W_y) = (\|B_x - XW_x\|_F^2 + \|B_y - YW_y\|_F^2) - \\ &\alpha' \sum_{(i,j)} S_{ij} \left(x_i^T W_x W_y^T y_j - \sqrt{x_i^T W_x W_x^T x_i} \sqrt{y_j^T W_y W_y^T y_j} \right) \\ &s.t.\, W_x^T W_x = I_{c\times c}, W_y^T W_y = I_{c\times c} \end{aligned} \tag{2.12}$$

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, Lu et al. have presented an efficient Discrete Latent Semantic Hashing (DLSH) approach [120]. 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 two modalities is given as:

$$\begin{aligned} &\min_{U_i|_{i=1,2}, A_i|_{i=1,2}, W_i|_{i=1,2}, Q} \sum_{i=1}^{2} \|\phi(X_i) - U_i A_i\|_F^2 + \\ &\beta \sum_{i=1}^{2} \|B - W_i A_i\|_F^2 + \\ &\delta \|B - QY\|_F^2 + \gamma \left(\sum_{i=1}^{2} \|U_i\|_F^2 + \sum_{i=1}^{2} \|W_i\|_F^2 + \|Q\|_F^2 \right) \\ &s.t. B \in \{-1, 1\}^{L\times N} \end{aligned} \tag{2.13}$$

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.

Shen et al. 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 [2]. 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 Figure 2.4.

Tang et al. have proposed an approach of supervised matrix factorization hashing for using label information and effective cross-modal retrieval [124]. 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, [115] 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.

Yu et al. utilizes multiple views for image and text representation to enhance feature information [123]. 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.

2.2.2 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 [113]. Thus, the multimedia 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 [121]. 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 [122] 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 [118]. 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 [127] 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 [119] 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.

Liu et al. introduced a novel cross-modal zero-shot hashing method which efficiently utilizes both labeled and un-labeled multi-modal data having separate label spaces [125]. 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. Ji et al. 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 [135]. It can capture the binary similarity as well as the complex multi-label semantic structure of data in diverse forms at the same time.

2.3 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. Summary of prominent multi-modal datasets is given in Table 2.2 which includes dataset name, mode, total concepts, dataset size, image representation, text representation, related article, and data source. Figure 2.5 presents a graph of the total number of categories in the datasets. Information regarding prominent benchmark datasets is given in the following points.

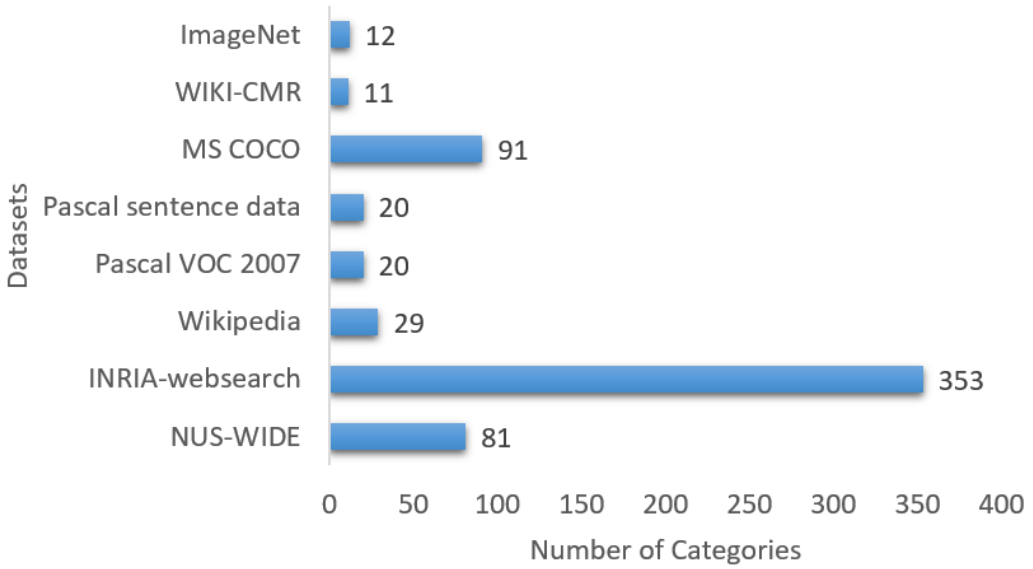


Figure 2.5: A chart displaying the total number of categories in the popular datasets

1. *NUS-WIDE*[1] [136]: 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 5,018 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,

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

64-D color histogram, 500-D BoW based on SIFT descriptions, 73-D edge direction histogram and 225-D block-wise color moments.

2. *IAPR TC-12*[1] [137]: 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 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*[2] [43]: 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 2,866 documents. These are image-text pairs that have been assigned a class from the vocabulary of 10 semantic classes. Figure 2.6 shows two examples (art and biology class) from the dataset with the image and the associated text (paragraph).

4. *PASCAL VOC 2007*[3] [138]: 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*[4] (photo and video sharing website). The dataset consists of a total of 9,963 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[5] for appropriately annotating each

[1] `https://www.imageclef.org/photodata`
[2] `http://www.svcl.ucsd.edu/projects/crossmodal/`
[3] `http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html`
[4] `https://www.flickr.com/`
[5] `http://host.robots.ox.ac.uk/pascal/VOC/voc2007/guidelines.html`

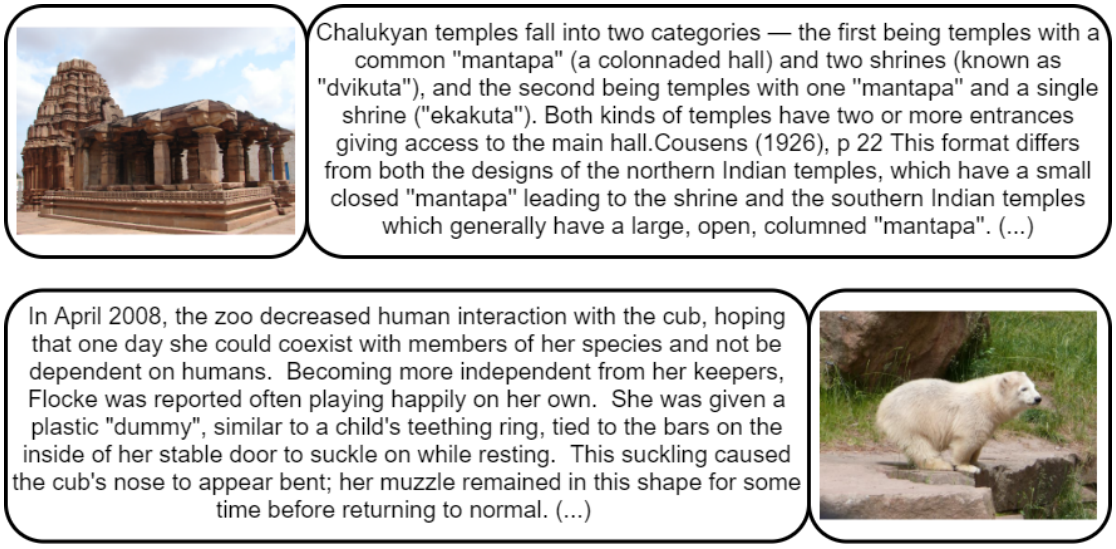


Figure 2.6: Two examples from Wikipedia dataset in which an image is associated with the collateral text

image in the ground-truth[138]. The entities mentioned in the annotation are class, bounding box, view, truncated, and difficult.

5. *MIR Flickr 25k and 1M*[1] [139], [140]: 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* [141]: 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.

[1] `http://press.liacs.nl/mirflickr/`

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 × 150 pixel square, however, preserving the original aspect ratio.

7. *Flickr 8k and 30k*[1] [142], [143]: Flickr 30k is an extension of the Flickr 8k dataset. Both datasets have been created from the Flickr website. Flickr 8k contains 8,092 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 [142].

8. *PASCAL Sentence Data*[2] [144]: The images for this dataset have been collected from PASCAL VOC 2008 challenge [138]. 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*[3] [145]: 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 labelled using per-instance segmentation to help in precise object localization. The dataset consists of total 3,28,000 images with 25,00,000 labelled instances. The objects chosen for the dataset are from 91 diverse categories. The annotation pipeline has been divided into three prominent exercises: (1) labelling concepts which are present in the image, (2) locating and marking all instances of labelled concepts; and (3) segmentation of each object instance.

10. *WIKI-CMR* [49]: 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 11 diverse

[1] `http://shannon.cs.illinois.edu/DenotationGraph/`
[2] `http://vision.cs.uiuc.edu/pascal-sentences/`
[3] `http://cocodataset.org/`

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.

Table 2.2: Summary of prominent image-text multi-modal datasets

Sr	Dataset	Year	Mode	Total concepts	Total images/ text	Image representation	Text representation
1	IAPR TC-12 [137]	2006	Image/ caption	Diverse	20,000/ 60,000	-	-
2	MIRFlickr 25k [139]	2008	Image/ tags	Diverse	25,000/ 2,23,500	-	-
3	NUS-WIDE [136]	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[1] [146]	2009	Images/ synsets	12 subtrees	32,00,000/ 5,247	SIFT	-
5	Wikipedia [43]	2010	Image/ text	29 (10 major)	2,866/ 2,866	SIFT	LDA
6	Pascal VOC 2007 [138]	2010	Image/ tags	20	9,963/ 24,640	-	-
7	MIRFlickr 1M [140]	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 [141]	2010	Image/ labels	353	71478/ -	-	-
9	Pascal sentence data [144]	2010	Image/ sentences	20	1000/ 5000	-	-
10	Wikipedia POTD [147]	2011	Images/ paragraphs	NA	1987/ 1987	SIFT	Text tokenization using rainbow
11	Flickr 8k [142]	2013	Image/ captions or sentences	Diverse	8092/ 40460	-	-
12	WIKI-CMR [49]	2013	Images/ paragraphs/ hyperlinks	11	38,804/ 74,961	SIFT, gist, PHOG, LBP, self similarity, spatial pyramid method	TF-IDF
13	Flickr 30k [143]	2014	Image/ captions or sentences	Diverse	31,783/ 1,58,915	-	-
14	MS COCO [145]	2014	Images/ labels	91	3,28,000/ 25,00,000	-	-

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

2.4 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 [5]. 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 [73]. 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 [68, 78]. Refer to Table 2.3 for a complete understanding of the definition of precision and recall. Precision and recall can be expressed as (Eqs. 2.14 and 2.15):

$$precision = \frac{TP}{TP + FP} \tag{2.14}$$

$$recall = \frac{TP}{TP + FN} \tag{2.15}$$

   where TP indicates true positive, FP is false positive and FN represents false negative.

   Several researchers have used the precision-recall curve to visualize the performance of their algorithm [148, 1, 55, 149]. The curve indicates the precision value at different recall levels. Authors in [150] have also used precision curve for performance visualization. It indicates the change in precision with respect to the number of retrieved results.

Table 2.3: 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	$TP+FP+TN+FN$

2. *F-measure*: It is a typical metric utilized for evaluating the performance of information retrieval systems [78]. After considering the effects of both precision and recall, F-measure can be defined mathematically as eq. (2.16):

$$F = \frac{(\theta^2 + 1) \times precision \times recall}{\theta^2 \times (precision + recall)} \tag{2.16}$$

here θ has been used for adjusting the weighted proportion of both recall and precision. If θ becomes 1 then F-measure can be redefined as F_1 (Eq. 2.17):

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{2.17}$$

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) [73]. 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. 2.18):

$$AP = \frac{1}{R} \sum_{o=1}^{O} P(o) rel(o) \tag{2.18}$$

where R is the number of relevant outcomes in the retrieved outcomes, $P(o)$ is the precision of top o retrieved outcomes, if the o^{th} retrieved outcome is relevant then $rel(o) = 1$ and otherwise 0. Now, MAP can be defined as (Eq.

2.19):

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP \tag{2.19}$$

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 [78, 73, 147] 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 [78, 73] for method evaluation regarding the position of the corresponding ground-truth outcome paired with the query. It is mathematically expressed as (Eq. 2.20):

$$MRR = \frac{1}{|O|}\sum_{n=0}^{|O|} \frac{1}{rank_n} \tag{2.20}$$

where $|O|$ is the number of query outcomes, $rank_n$ indicates the position of corresponding unique groud-truth paired with n^{th} query in the retrieved set.

2.5 Discussion and comparative analysis

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 2.6 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

Table 2.4: Comparison of hashing techniques in diverse supervision modes

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

Table 2.5: 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

deep learning. The table describes the cross-modal method, image and text feature extractors, the dataset used, method type, evaluation metric, and references.

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 2.4 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 multimodal 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 [114]. The refinement and potent of feature extraction part of the deep hashing method must be considered in future works. Table 2.5 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 2.6: Summary of works done in image-text cross-modal retrieval

SR.	**TECHNIQUE**	**IMAGE REP.**	**TEXT REP.**	**DATA**	**TYPE**	**METRIC**	**REF. & YR.**
Real-valued representation techniques							
1	Linkage of each image feature with text feature	Bag-of-Words	CiCui system [151], TF-IDF and weirdness [152]	US FEMA flood data, Facebook pages' and groups' data related to floods	-	MAP	[37] 2016
2	Structural SVM, SR, CSR (using CCA)	BoW of dense SIFT features	Probability distribution	UIUC Pascal Sentence dataset, IAPR TC-12 benchmark and SBUCaptioned Photo dataset	-	BLEU score, rouge score	[45] 2014
3	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 [49] dataset	Semi-supervised	MAP	[61] 2015
4	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	[102] 2011
5	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	[68] 2015

6	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	[39] 2014
7	Joint feature selection and subspace learning	Gist, SIFT	LDA	Pascal, Wikipedia and NUS-WIDE dataset	-	MAP, PR curve	[53] 2015
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	[153] 2015
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	[52] 2012
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	[46] 2017

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	[154] 2018
12	CM, SM, SCM	SIFT	LDA	Wikipedia	Unsupervised	MAP, PR curve	[43] 2010
13	Graph regularization and modality dependence (GRMD)	CNN	LDA	INRIA-websearch, Pascal sentence, Wikipedia 2010	-	MAP, PR curve	[1] 2020
14	CCA, KCCA	SIFT, gist, PHOG, LBP, self similarity, spatial pyramid method	TF-IDF	WIKI-CMR	-	Precision	[49] 2013
15	Improved CCA	-	-	NUS-WIDE, Pascal sentence, Wikipedia	-	MAP	[48] 2017
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)	[51] 2010
17	MLRank	Gist, color histogram, color sutocorrelation, edge direction histogram, wavelet texture, blockwise color moments	-	Corel 5k, NUS-WIDE, IAPR TC12	Semi-supervised	Precision, recall, F1 score, MAP, N+ (no. of keywords with non-zero recall value)	[69] 2013
18	CM, SM, SCM	SIFT	LDA	TVGraz, Wikipedia	Supervised and unsupervised	MAP, PR curve	[44] 2013
19	CCA and its probabilistic interpretation	RGB-SIFT	Binary features	MIRFlickr 1M	-	Precision	[47] 2014
20	Regularizer of image semantics	SIFT	LDA	TVGraz, Wikipedia, pascal sentence dataset	-	MAP, PR curve	[103] 2014

21	Modality-dependent cross-media retrieval (MDCR) model	CNN visual features	LDA	Wikipedia, pascal sentence, INRIA-websearch	Supervised	MAP	[59] 2016
22	Semantic consistency cross-modal retrieval (SCCMR)	CNN, VGG	LDA, BoW	Wikipedia, pascal sentence, NUS-WIDE-10k, INRIA-websearch	Semi-supervised	MAP, PR curve	[54] 2020
Binary-valued or cross-modal hashing techniques							
23	Unsupervised Concatenation Hashing (UCH)	Gist	word frequency count	Pascal, UCI handwritten digit, Wikipedia	Unsupervised	MAP	[126] 2019
24	Cross-modal self-taught hashing (CMSTH)	SIFT, HoG, GIST	TF-IDF	Wikipedia, NUS-WIDE	Unsupervised	MAP	[113] 2016
25	Linear cross-modal hashing	SIFT	LDA	NUS-WIDE, Wikipedia	-	MAP, recall	[116] 2013
26	Latent semantic sparse hashing	Sparse coding	Matrix factorization	Labelme, NUS-WIDE, Wikipedia	-	MAP, PR curve	[134] 2014
27	Quantized correlation hashing	SIFT, BoW	LDA, tag vector	58W-CIFAR, NUS-WIDE, Wikipedia	Supervised	MAP, precision	[117] 2015
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, MIR-Flickr 25k, NUS-WIDE, Wikipedia	Supervised	MAP, PR curve	[120] 2019
29	Subspace Relation Learning for Cross-modal Hashing	SIFT, gist	LDA, tag occurrence feature vector	ImageNet, Labelme, MIR-Flickr 25k, NUS-WIDE, UCI handwritten digit data, Wikipedia	Supervised	MAP, precision	[2] 2020
30	Deep Visual Semantic Hashing model	Deep $fc7$ features	BoW vector	IAPR TC-12, MS COCO	-	MAP, precision	[121] 2016
31	Deep cross-modal hashing	Gist, bag-of-visual-words (BOVW)	BoW vector	IAPR TC-12, MIRFlickr 25k, NUS-WIDE	-	MAP, PR curve	[122] 2017

32	Triplet-based deep hashing network	SIFT	BoW	MIRFlickr 25k, NUS-WIDE	Supervised	MAP, PR curve	[118] 2018
33	Attention-Aware Deep Adversarial Hashing	-	BoW	IAPR TC-12, NUS-WIDE, MIRFlickr 25k	-	MAP, PR curve	[127] 2018
34	Supervised matrix factorization hashing	SIFT	Topic vector, BoW	NUS-WIDE, Wikipedia	Supervised	MAP, precision, PR curve	[124] 2016
35	Semantic deep cross-modal hashing	-	BoW	IAPR TC-12, MIRFlickr, NUS-wIDE	Supervised	MAP, precision curve, PR curve	[119] 2019
36	Zero-shot hashing	BoVW, SIFT, gist	LDA, BoW	MIRFlickr, NUS-WIDE, Wikipedia	Semi-supervised	MAP	[125] 2019
37	Deep multi-level semantic hashing	-	BoW	MIRFlickr 25k	Supervised	MAP, PR curve	[135] 2019
38	Cycle-Consistent Deep Generative Hashing (CYC-DGH)	CNN	LDA	Microsoft COCO, IAPR TC-12, wiki	-	MAP, precision curve, PR curve	[155] 2018
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	[115] 2019
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	[123] 2020
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	[76] 2012
42	Levinberg-Marquardt deep canonical correlation analysis (LMDCCA)	Deep neural network	Deep neural network	Wikipedia Articles data	-	Precision recovery curve	[77] 2019

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	[156] 2016
44	Deep canonical correlation analysis(DCCA)	color histogram, color correlogram, edge direction histogram, wavelet texture, block-wise color moments, SIFT, GIST, MPEG-7	Bag-of-Words (BoW)	Wikipedia, pascal, NUS-WIDE10k	Supervised	MAP	[157] 2016
45	Deep coupled metric learning (DCML) method	SIFT, BoVW, GIST, color histogram	Latent Dirichlet allocation (LDA) model	Wikipedia, Pascal VOC 2007, NUS-WIDE	-	Precision, MAP, ROC and CMC curve	[158] 2016
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	[81] 2019
47	Correspondence autoencoder	Pyramid Histogram of Words (PHOW), MPEG-7 descriptors and Gist	Bag-of-Words	Wikipedia, Pascal, NUS-WIDE	-	MAP	[80] 2015
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	[159] 2019

49	Deep Adversarial Metric Learning approach (DAML)	SIFT, VGG	LDA, BoW	Wikipedia, pascal, NUS-WIDE	Supervised	MAP	[91] 2019
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	[160] 2019
51	Deep Belief Network	LBP	-	NUS-WIDE, Wikipedia	-	MAP, percentage, MRR	[78] 2017
52	Log-Bilinear Model	-	-	IAPR TC-12, attribute discovery, SBU captioned photos	-	Bleu, perplexity and retrieval evaluation	[82] 2014

2.6 Research gaps

Based on the literature review, the research gaps which have been identified are:

1. *Noisy and restricted annotations*: An enormous amount of multi-modal data is composed by researchers from various websites such as Flickr, YouTube, and Facebook to name a few. This data on the web is mostly unstructured and the accompanying annotations are also noisy and restricted. Annotations provide the semantic information required to understand an image or a video and labeling a large number of instances manually is almost impossible. In [161], 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.

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 [162, 163, 164]. Although,

miscellaneous authors have utilized these methods for cross-modal system design, however, they are still in their early stage and require to be explored more. Moreover, researchers either use a soft-computing or algorithmic approach. Both have their strengths and weaknesses. So, there is a need for a hybrid approach to integrate heterogeneous data of diverse modalities.

3. *Lack of large scale multi-modal datasets*: Researchers are proposing different algorithms nowadays for cross-modal retrieval and annotation. However, there is a lack of colossal datasets comprising various modalities to test and validate the introduced algorithms. In most of the articles, 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 need to compose large scale multi-modal datasets and especially in the medical field [165].

4. *Confusion in choosing data feature extraction method*: Most of the techniques introduced by authors consist of individual feature extraction from each modality to be used in cross-modal framework design as a first step. If the initial step is inappropriate then it will affect the implementation of the whole framework. For instance, the performance of a machine learning model extremely depends upon the feature representation used for building a model. This happens because different feature representations hide more or less diverse descriptive factors of variations behind the data [166]. So, it is necessary to choose a suitable 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 [167]. Moreover, further research is required for designing efficient cross-modal algorithms which can be tested on huge multi-modal datasets [168].

6. *Need of novel and diverse fields' datasets*: It has been identified from the literature that most of the datasets are comprised of images and captions from social media websites, so their content is quite 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 [123]. However, most of the generated multi-media data is either 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 [54].

2.7 Objectives

The research objectives are as follows:

1. To study, analyze and explore several existing methods available for multi-modal data fusion and identify important aspects which needs consideration.

2. To propose and implement cross-modal framework to incorporate rich textual information with images stored in repository to enhance the information retrieval.

3. To test and validate the proposed framework using quantitative analysis on various public and primary datasets.

Chapter 3
Hybrid Self-organizing Map Approach

3.1 Overview

The following points provide the overview of the chapter:

1. The proposed hybrid SOM (HSOM) method integrates the image and text modalities and ensures an effective cross-modal retrieval process.

2. Two SOMs are trained separately using image and collateral text feature vectors.

3. The presented HSOM framework is influenced by the working of the brain as diverse data representations (in the form of two separately trained SOMs) are combined easily using a Hebbian network.

A concept of information fusion comprising image and text is studied in this chapter which is inspired by the working of the brain. In the proposed cross-modal retrieval framework, two Self-Organizing Maps (SOM) are trained independently for images and collateral text and then fused using the Hebbian network. The introduced algorithm can be applied to construct systems that can learn to integrate diverse data modalities (images and text in our case). The framework comprises: (1) a SOM trained to cluster images; (2) another SOM learns the text; and (3) the Hebbian network links the highly active nodes on image SOM with nodes on text SOM. The final system after merging both image and text SOM using Hebbian network is known as hybrid SOM (HSOM). Figure 3.1 demonstrates the flow diagram of the proposed cross-modal HSOM framework.

3.2 Modalities' feature vector creation

Feature vectors have been extracted from each image and text in the dataset as described in the following subsections. The goal is to utilize the prevalent image analysis method for creating vectors that can define the diverse and significant properties of an image such as color, texture, and edges. Zernike moments (ZM) have been considered here for image vector creation because of their efficacy in extracting prominent image features [169, 170, 171]. Latent Dirichlet Allocation

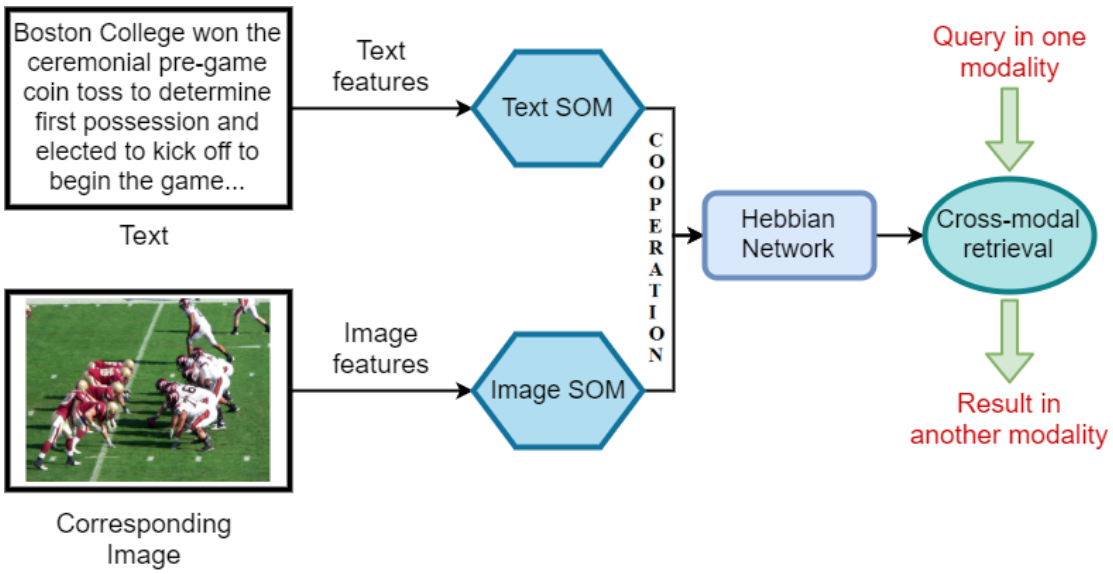


Figure 3.1: Flow diagram of the proposed cross-modal information retrieval system

(LDA) [172] model is utilized for extracting text features due to its prominence in text analytic area [173, 174]. Along with the combination of ZM and LDA feature transforms for the modalities' representation, Scale Invariant Feature Transform (SIFT) features have also been utilized for visual feature representation accompanied by LDA for text in the experimentation. These features are available with the dataset and used by numerous researchers for comparison purposes.

3.2.1 Image features

Image features capture shape, color, and texture values based upon the given dimension (details) and discontinuities in the image. In this study, Zernike moments (ZM) have been extracted from images as their features to represent them. Figure 3.2 shows the steps followed by each image for ZM extraction which can distinguish it from other images in the data. Each image is pre-processed before calculation of ZM which includes image resizing to 1000×1000 size, RGB to grayscale conversion (if it is not grayscale), and image normalization [175]. The steps followed for image pre-processing are described as follows:

Image pre-processing

Image pre-processing is a very crucial step of image recognition process as it increases the accuracy and make the proposed technique more efficient. The aim of pre-processing is to improve the image data, getting better feature extraction, gaining consistency and also improves some image features important for additional processing. In this study, three steps are performed under pre-processing

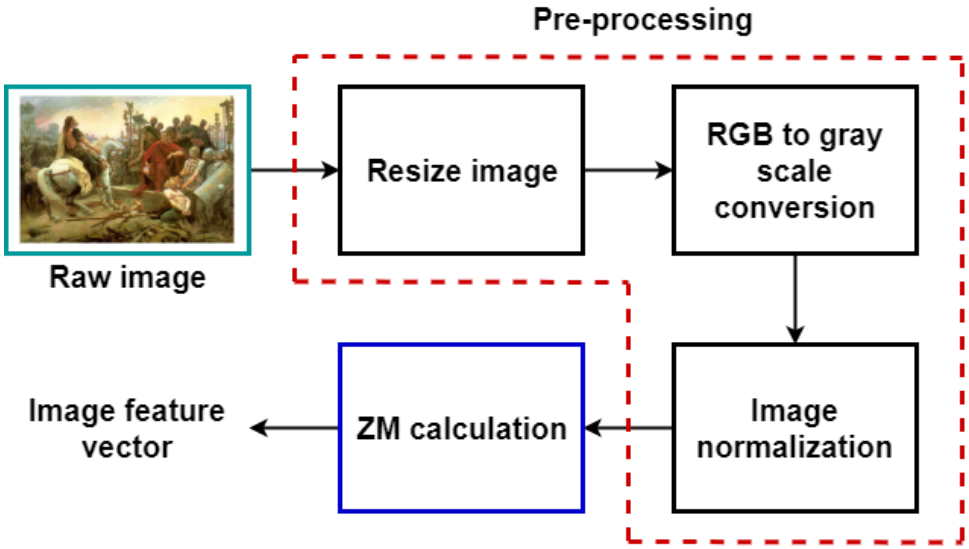


Figure 3.2: Process followed by each image for ZM extraction

stage: (a) image resizing, (b) conversion of image from RGB to gray scale, and (c) normalization.

- **Resize**: The collected images are all of different sizes having different number of pixels. So, in order to get accurate results, image resizing has an important role. All the images in the dataset are resized to 1000×1000 size which make them square in shape and of same size.

- **Conversion**: Conversion of images from RGB scale to gray scale is also necessary as it increases speed of processing, decreases code complexity and gray images can be easily visualized.

- **Normalization**: Normalization is vital to ensure that each pixel has a similar data distribution. It calibrates the diverse pixel intensities into a normal distribution and makes the computation efficient by updating all the pixel values between 0 to 1. Let Min and Max are minimum and maximum intensity values of image and $newMin$ and $newMax$ are the new intensity values, then the linear normalization (I_n) of a gray scale image can be calculated as:

$$I_n = (I - Min)\frac{newMax - newMin}{Max - Min} + newMin \tag{3.1}$$

Zernike Moments (ZM)

Moments are the projections of image functions onto particular kernel functions. They represent the weighted average of the intensity values of image pixels to

obtain the scalar quantities for image interpretation. They can be defined in Polar and Cartesian coordinate system. $G_{nm}(x, y)$ denotes the basis or kernel function. With change in kernel functions, different types of moments can be obtained. Now, these moments are said to be orthogonal if they satisfy the following condition:

$$\int_0^1 \int_0^1 G_{nm}(x, y) G_{st}(x, y) dxdy = k\delta_{ns}\delta_{mt} \tag{3.2}$$

here k is a normalization coefficient and δ_{ns}, δ_{mt} represents Kronecker delta which is expressed as:

$$\delta_{ns} = \begin{cases} 1, & n = s \\ 0, & otherwise \end{cases} \tag{3.3}$$

If the moments do not satisfy above condition in Equation 3.2 then they are called non-orthogonal. Non-orthogonal moments include complex moments (with kernel function $G_{nm}(x, y) = (x + iy)^n (x - iy)^m$), geometric moments ($G_{nm}(x, y) = x^n y^m$) and rotational moments ($G_{nm}(r, \theta) = r^n e^{-im\theta}$) which are defined in [176, 177]. Non-orthogonal moments face difficulties in image reconstruction due to lack of orthogonal nature. They are highly noise prone and information redundant. However, orthogonal moments are rotation, scaling and translation invariant, robust to image noise and possess minimal information redundancy [178]. Due to these advantages, they have been used in various digital image processing areas such as image clustering [179], brain MRI segmentation [180], handwritten numeral recognition [181], and face recognition [182].

ZM are a type of continuous orthogonal moments. Moments of different order provide varying information about the image, such as the center of mass, area, intensity, and orientation. Comparative to other existing orthogonal moments, ZM are chosen as they provide valuable results in image representation and need lower computational precision for this task. ZM are devised from complex Zernike polynomials which were introduced by an optical physicist named *Frits Zernike* [183]. These are a series of polynomials defined within a unit circle over the space of polar coordinates. Fundamentally, ZM are the projections of an image function along real and imaginary axes (x- and y-axis) which is convolved by an orthogonal function. Hence, they represent images in different frequency components such as orders (along radial direction) and repetitions (along angular direction). Figure 3.3 demonstrates the first 27 Zernike polynomials starting from order 1 where n is order of moment and m is multiplicity. They are vertically ordered as per radial degree and horizontally according to azimuthal degree. Azimuth is an angle between a reference vector and the projected vector on the reference plane. In ZM, an image is mapped onto a unit circular disk such that the center of the

image is transformed into the center of the disk. This mapping can be performed in two ways: (1) inner circle mapping (Figure 3.4); and (2) outer circle mapping (Figure 3.5) [184]. In the former, corner image pixels are excluded while computing moments which results in information loss and is a drawback especially when corners are informative. Therefore, a perfect square to circular domain mapping cannot be obtained and the circular boundary is approximated in a zig-zag pattern (Figure 3.4(b)). However, the complete image is mapped onto the disk in outer circle mapping avoiding any information loss. Due to this advantage, outer circle mapping has been utilized while computing ZM in this study.

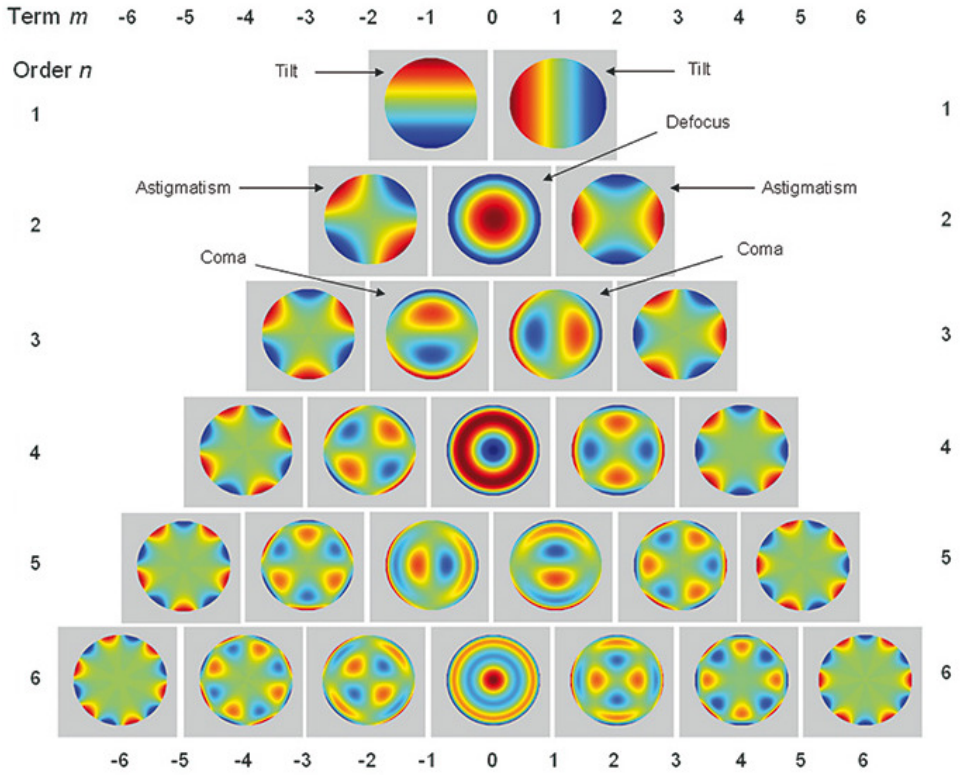


Figure 3.3: First 27 Zernike polynomials

According to [185], if $f(r, \mu)$ depicts an image function, then the two dimensional ZM with order s and repetition t can be defined in Polar coordinate system as:

$$Z_{st} = \frac{s+1}{\pi} \int_0^{2\pi} \int_0^1 f(r, \mu) V_{st}^*(r, \mu) r dr d\mu \tag{3.4}$$

here $V_{st}^*(r, \mu)$ represents the complex conjugate of zernike polynomials depicted as V_{st} and is defined as:

$$V_{st}(r, \mu) = R_{st}(r) e^{it\mu} \tag{3.5}$$

which satisfies $s \geq 0$, $0 \leq \mid t \mid \leq s$, $s - \mid t \mid = \text{even}$, $i = \sqrt{-1}$ and $\mu =$

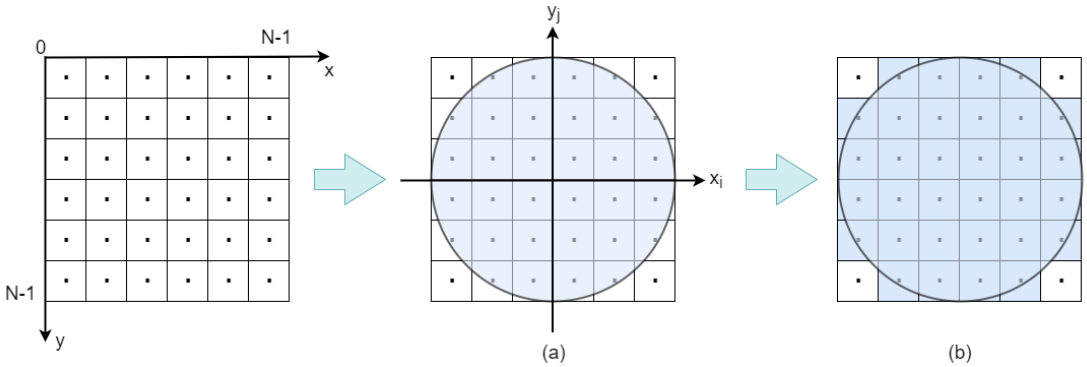


Figure 3.4: Inner circle mapping technique. (a) image mapping onto unit disk; (b) inscribed disk approximated by square grids

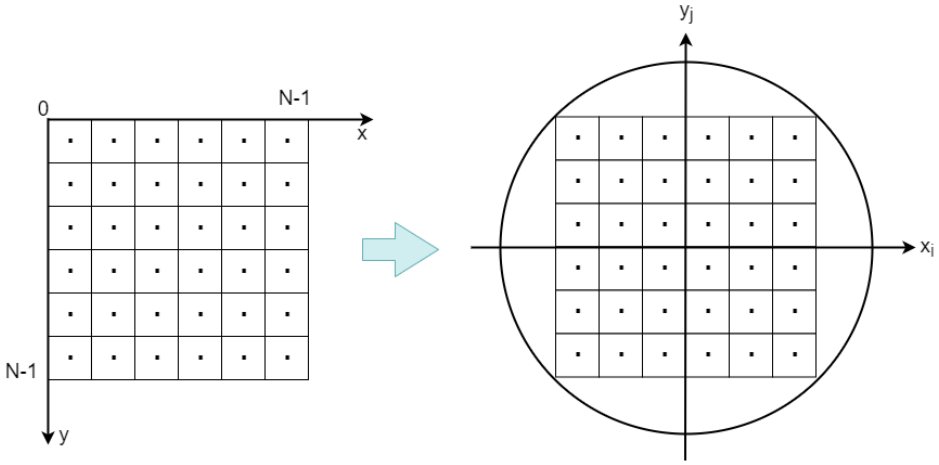


Figure 3.5: Outer circle mapping technique where complete image is mapped inside the disk

$\arctan(y/x)$. (r, μ) are radius and angle of pixel from origin, which means polar coordinate of a pixel at (x, y). Radial Polynomials are given as:

$$R_{st}(r) = \sum_{k=0}^{(s-|t|)/2} (-1)^k \times \frac{(s-k)!}{k!(\frac{s+|t|}{2}-k)!(\frac{s-|t|}{2}-k)!} r^{s-2k} \tag{3.6}$$

The integrals in Equation 3.4 are replaced by summations for a digital image. So, Equation 3.4 becomes:

$$Z_{st} = \frac{s+1}{\pi} \sum_x \sum_y f(x,y) V_{st}^*(x,y) \Delta x \Delta y \qquad , x^2 + y^2 \leq 1 \tag{3.7}$$

If $f(x, y)$ is a digital image of size $N \times N$ [184], then ZM is given as:

$$Z_{st} = \frac{s+1}{\pi} \sum_0^{N-1} \sum_0^{N-1} f(x_i, y_i) V_{st}^*(x_i, y_i) \Delta x_i \Delta y_i \tag{3.8}$$

With center of pixel (i, j), (x_i, y_i) are the mapped pixel coordinates that occupies the area of $[x_i - \Delta x/2, x_i + \Delta x/2] \times [y_j - \Delta y/2, y_j + \Delta y/2]$ and are defined as equations (3.9) and (3.10):

$$x_i = \frac{2i + 1 - N}{D} \tag{3.9}$$

$$y_j = \frac{2j + 1 - N}{D} \tag{3.10}$$

where Δx and Δy can be expressed as:

$$\Delta x_i = \Delta y_j = \frac{2}{D} \tag{3.11}$$

Using equations (3.9, 3.10) and (3.11), Eq. 3.8 can be written as:

$$Z_{st} = \frac{4(p+1)}{\pi D^2} \sum_{0}^{N-1} \sum_{0}^{N-1} f(x_i, y_i) V_{st}^*(x_i, y_i) \tag{3.12}$$

Here, the value of D should be chosen as per the mapping technique. For inner circle mapping (Figure 3.4), the value of D will be N, and for outer circle mapping approach (Figure 3.5), it will be $N\sqrt{2}$. ZM can be calculated using Equation 3.12 by replacing D with N, in case of inner circle mapping. However, in outer circle mapping that is implemented in the proposed study, ZM can be calculated as:

$$Z_{st} = \frac{2(p+1)}{\pi N^2} \sum_{0}^{N-1} \sum_{0}^{N-1} f(x_i, y_i) V_{st}^*(x_i, y_i) \tag{3.13}$$

Rotation and scale invariance can be obtained in ZM by normalizing the image via Cartesian moments before ZM calculation. Translation invariance can be obtained if image's centre of mass is shifted to origin [186]. The obtained number of moments (NoM) according to given order s can be evaluated as:

$$NoM = \begin{cases} \frac{1}{4}(s+1)(s+3), & s = odd \\ \frac{1}{4}(s+2)^2, & s = even \end{cases} \tag{3.14}$$

Scale invariant feature transform (SIFT)

SIFT features detect and describe the local features of a digital image. They are proposed by a Canadian computer scientist *David G. Lowe*. SIFT features represent an image using an extensive collection of local feature vectors, invariant to image rotation, scaling, translation, and partially invariant to illumination and

affine or 3D projection [187]. The primary objective behind introducing SIFT features was the lack of image features invariant to scaling and sensitivity towards illumination changes. The idea of SIFT features is inspired by the responses of neurons in the inferior temporal cortex in primary vision. The effective way of SIFT feature extraction is *cascade filtering method*, where the computationally expensive tasks are only applied at locations that pass an initial exam [188]. The prominent stages of computing SIFT features of a digital image are defined below:

1. *Scale-space extrema detection*: The first task in keypoint detection is to recognize the scales and locations assigned recurrently under different views of the same object. For detecting locations invariant to scale update of an image, stable features need to be searched across all potential scales, utilizing a continuous function of scale called scale-space. The Gaussian function has been used as a scale-space kernel. Thus, the scale-space of an image $I(x, y)$ is described as a function $L(x, y, \sigma)$, which is created from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{3.15}$$

here '*' represents the convolution operation in x and y.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{3.16}$$

For detection of stable keypoints, scale-space extrema in the difference-of-Gaussian (DoG) function $D(x, y, \sigma)$ convolved with the image, has been employed. It can be evaluated using the difference of two nearby scales partitioned by a constant multiplicative factor p:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, p\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, p\sigma) - L(x, y, \sigma) \end{aligned} \tag{3.17}$$

Figure 3.6 represents an effective technique for constructing $D(x, y, \sigma)$. For each octave of scale-space, the initial image is convolved with Gaussians recurrently to generate the set of scale-space images as depicted on the left. Adjacent Gaussian images are subtracted to create the difference-of-Gaussian images on the right. The Gaussian image is down-sampled by a factor of 2 after each octave, and the process is repeated. The accuracy of sampling relative to σ is almost similar to the start of the previous octave, while the computation is highly reduced. Each sample point is compared to

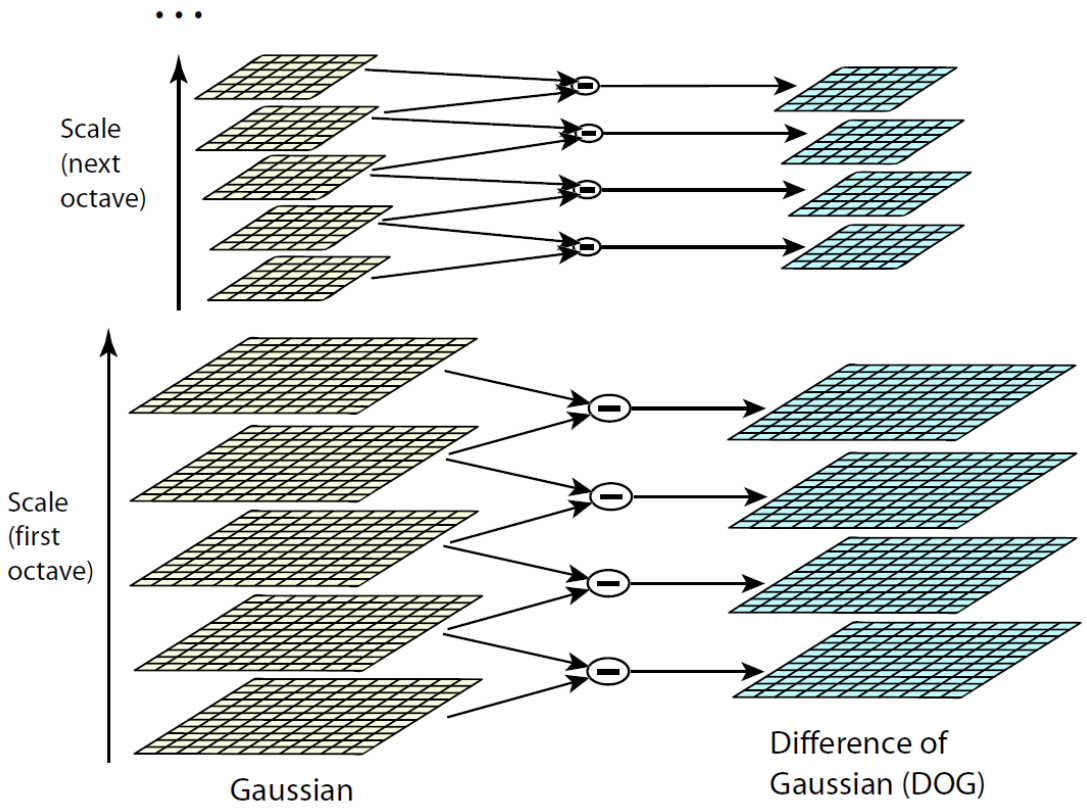


Figure 3.6: For each octave of scale-space, the initial image is convolved with Gaussians recurrently to generate the set of scale-space images as depicted on the left. Adjacent Gaussian images are subtracted to create the difference-of-Gaussian images on the right. The Gaussian image is down-sampled by a factor of 2 after each octave and the process is repeated. [188]

its eight neighbors in the current image and nine neighbors in the above and below scale to identify the local minima and maxima of $D(x, y, \sigma)$. This has been illustrated in Figure 3.7. The sample point is chosen only if it is larger or smaller than all the neighbors. The cost of checking this is considerably low because most of the sample points are already eradicated in first few checks.

2. *Keypoint localization*: After finding the keypoint candidate, detailed fit is performed to the nearby data for scale, location, and ratio of principal curvatures. It will help in rejection of the points having low contrast (noise sensitive) or the points which are poorly localized along the edge. For finding the interpolation of each keypoint, the scale-space DoG function $D(x, y, \sigma)$ can be expressed in a small 3D neighborhood around a keypoint using second-order Taylor series:

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x} \tag{3.18}$$

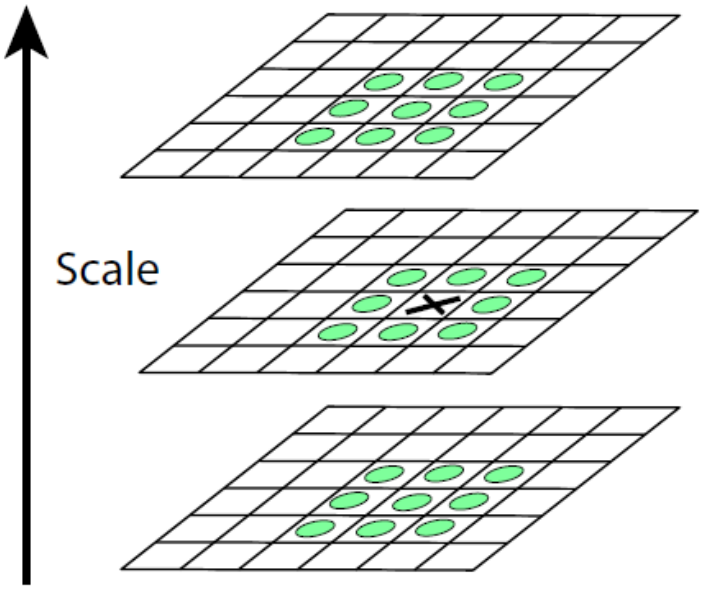


Figure 3.7: Maxima and minima of the DoG images are identified by comparing a pixel (marked with X) to its 264 neighbors in 3×3 regions at the current and adjacent scales (marked with circles) [188].

D and its derivatives are calculated at the sample point and $\mathbf{x} = (x, y, \sigma^T)$ depicts the offset from this point. To evaluate the extremum $\hat{\mathbf{x}}$ location, take derivative of this function w.r.t. $\mathbf{x}$ and set it to zero, that gives

$$\hat{\mathbf{x}} = -\frac{\partial^2 D^{-1}}{\partial \mathbf{x}^2}\frac{\partial D}{\partial \mathbf{x}} \tag{3.19}$$

Add the final offset $\hat{\mathbf{x}}$ to its sample point's location for retrieving interpolated estimate for the extremum location. For declining unstable extrema with low contrast, $D(\hat{\mathbf{x}})$ (function value at the extremum) is used. It can be achieved by substituting Equation 3.19 into (3.18), that gives:

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D^T}{\partial \mathbf{x}}\hat{\mathbf{x}} \tag{3.20}$$

3. *Orientation assignment*: Image rotation invariance can be obtained by allocating a congruous orientation to each keypoint as per the local image characteristics, and it represents the keypoint descriptor relative to this orientation. All evaluations can be performed in a scale-invariant way by using the scale of the keypoint to choose the Gaussian smoothed image L with the closest scale. For every image sample $L(x, y)$, at this scale, the orientation $\theta(x, y)$ and gradient magnitude $m(x, y)$ is predetermined using pixel

differences:

$$\theta(x,y) = \arctan\left(\frac{(L(x,y+1) - L(x,y-1))}{(L(x+1,y) - L(x-1,y))}\right) \tag{3.21}$$

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{3.22}$$

The gradient orientations of the sample points within a region around the keypoint, generates an orientation histogram.

4. *Keypoint descriptor*: Image location, scale, and orientation can be assigned to each keypoint by the previous steps. Now a descriptor for the local image region is determined, that is highly distinguishing and invariant to illumination and 3D viewpoint changes. The evaluation of the keypoint descriptor has been demonstrated in Figure 3.8. First step is sampling of the image gradient magnitudes and orientations around the keypoint location with the use of scale of keypoint for selecting the Gaussian blur level for the image. The descriptor and gradient orientations' coordinates are rotated w.r.t. keypoint orientation, to obtain orientation invariance. The gradients are predetermined for all levels of the pyramid to enhance efficacy. These are represented using small arrows at each location of the sample on left side of Figure 3.8.

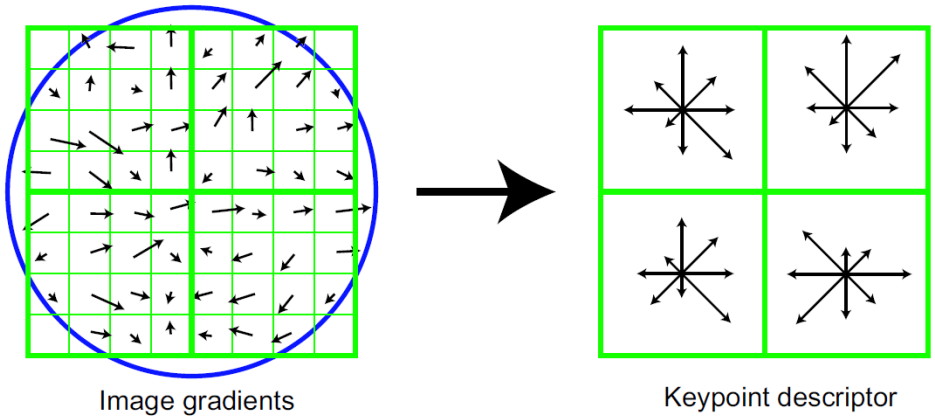


Figure 3.8: A keypoint descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the keypoint location, as shown on the left. These are weighted by a Gaussian window, indicated by the overlaid circle. These samples are then accumulated into orientation histograms summarizing the contents over 4×4 subregions, as shown on the right, with the length of each arrow corresponding to the sum of the gradient magnitudes near that direction within the region. This figure shows a 2×2 descriptor array computed from 8×8 set of samples. [188]

A weight is assigned to each sample point magnitude using the Gaussian weighting function with the value of σ equal half the size of the descriptor window width. It has been depicted using a circular window in Figure 3.8. The Gaussian window has been utilized for avoiding the abrupt changes in the descriptor with a minor change in window position and for giving the least importance to the gradients which are farther away from the descriptor center because of misregistration errors often infect these. The keypoint descriptor displayed on the right side of the Figure 3.8 creates orientation histograms over 4×4 sample regions, thus permitting a substantial shift in gradient positions. Eight directions have been illustrated in the figure for each orientation histogram. Each arrow's length depicts the magnitude of the histogram entry. The descriptor is created using a vector comprising all orientation histogram entries' values corresponding to the length of arrows. Figure 3.8 demonstrates the 2×2 array of orientation histograms; however, the 4×4 histograms array with 8 orientation bins has been applied in the experimentation. Hence, this study utilizes a $4 \times 4 \times 8 = 128$ element feature vector. In the end, the feature vector is updated to lessen the effects of change in illumination.

3.2.2 Text features

The dataset chosen for experimentation comprises image-text pairs such that for each image there is a corresponding textual paragraph(s) explaining it. It is necessary to choose the most important words from that text which can uniquely identify it. Text usually contains words like "a", "an" and "the" (known as stop words) in the highest numbers which are inessential for distinguishing a document from other documents. So, the text is pre-processed before the actual calculation of features. Latent Dirichlet Allocation (LDA) is one of the famous techniques for topic modeling which has been utilized here to extract the text features. Figure 3.9 shows the process of text feature matrix creation from a set of XML files. Firstly, the collateral text is extracted from each XML file and added into a string array. The collected strings in the array are pre-processed by decoding the HTML entities and removing tags, URLs, and numbers. Afterward, strings are converted into tokens and tokenized documents are created. These documents are pre-processed again which includes lemmatization of tokens, removal of punctuation marks, stop words and words having length 1 or 2. Then a bag of words is created from these cleaned documents which is further utilized to extract the LDA features.

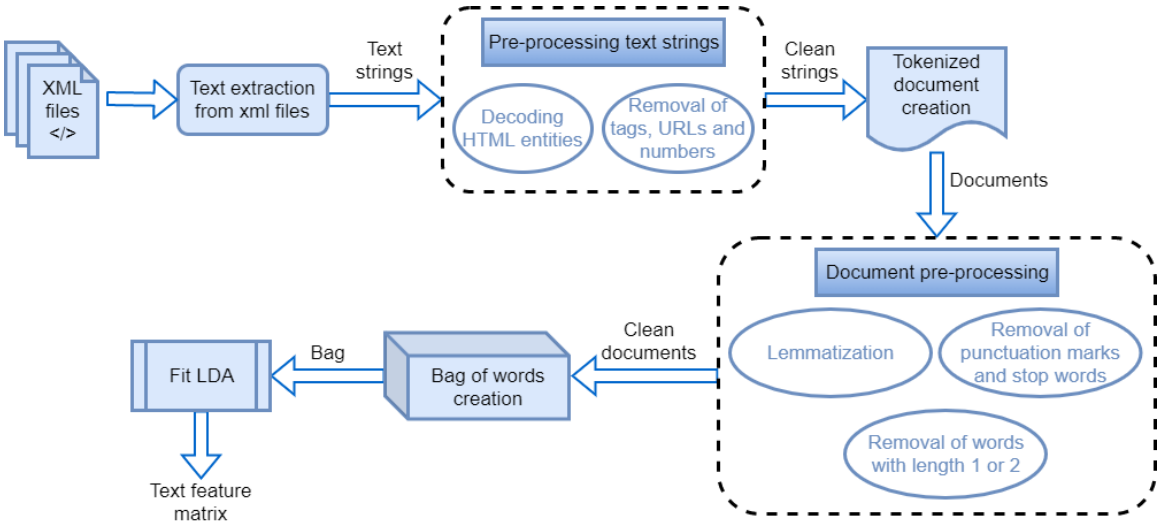


Figure 3.9: Process flow for text feature matrix creation

Latent Dirichlet Allocation (LDA)

Topic Modeling is a prominent technique in text mining and detecting relations among textual documents [189]. There are various methods for topic modeling, however, LDA is highly popular. It is a three-level hierarchical Bayesian model where documents are modeled as random finite mixtures over latent topics and each topic, in turn, is characterized as a word distribution [172]. A *word* can be described as a basic unit of discrete data and an element from vocabulary, a *document* refers to a series of R words designated by $\mathbf{w} = (w_1, w_2, ..., w_R)$ where w_r is r^{th} word in the sequence, and a group of Q documents indicated by $C = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_Q)$ is known as a *corpus*. Figure 3.10 represents LDA as a three-level probabilistic graphical model where the inner plate signifies the repetitive choice of topics and words in a document, however, outer plate denotes documents. γ and δ are parameters at the corpus level that are supposed to be sampled one time during generation procedure of the corpus. The η symbol represents variables at the document level that are sampled once in a document, however, z and w signify variables at the word level that are sampled once in a document for a single word.

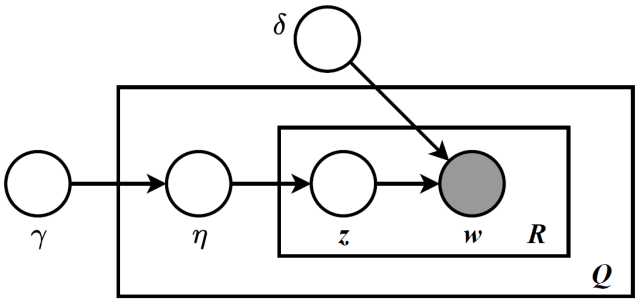


Figure 3.10: Graphical model representation of LDA

The generative procedure followed for each document $\mathbf{w}$ in a corpus C in LDA is given below [172]:

1. Choose $R \sim \text{Poisson}(\xi)$.
2. Choose $\eta \sim \text{Dir}(\gamma)$.
3. For each word w_r in a document, choose:
   (a) a topic $z_r \sim \text{Multinomial}(\eta)$.
   (b) a word w_r from $p(w_r|z_r, \delta)$, a multinomial probability conditioned on z_r topic.

Few assumptions which are made in the LDA basic model are: (1) dimensionality of Dirichlet distribution is well known and stable; (2) R is independent of other data creating variables such as η and $\mathbf{z}$; and (3) the word probabilities are parameterized by δ matrix where $\delta_{ij} = p(w^j = 1|z^i = 1)$, which is treated as a stable quantity that is to be evaluated.

A k-dimensional Dirichlet random variable η can have values in $(k-1)$-simplex (a k-vector η lies in the $(k-1)$-simplex if $\eta_i \geq 0$ and $\sum_{i=1}^{k} \eta_i = 1$) and has the below probability density on this simplex:

$$p(\eta|\gamma) = \frac{\Gamma(\sum_{i=1}^{k} \gamma_i)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \eta_1^{\gamma_1 - 1} ... \eta_k^{\gamma_k - 1} \tag{3.23}$$

where γ represents a k-vector with $\gamma_i > 0$ and $\Gamma(x)$ denotes Gamma function.

Given the parameters γ and δ, the joint distribution of a topic mixture η, a set of R topics $\mathbf{z}$, and a set of R words $\mathbf{w}$ can be defined as:

$$p(\eta, \mathbf{z}, \mathbf{w}|\gamma, \delta) = p(\eta|\gamma) \prod_{r=1}^{R} p(z_r|\eta) p(w_r|z_r, \delta) \tag{3.24}$$

here $p(z_r|\eta)$ is η_i for unique i such that $z_r^i = 1$. Integrating over η and summing over z, the marginal distribution of a document can be evaluated as follows:

$$p(\mathbf{w}|\gamma, \delta) = \int p(\eta|\gamma) \left(\prod_{r=1}^{R} \sum_{z_r} p(z_r|\eta) p(w_r|z_r, \delta) \right) d\eta \tag{3.25}$$

Finally, the probability of a corpus can be obtained by taking the product of the marginal probabilities of single documents:

$$p(C|\gamma, \delta) = \prod_{c=1}^{Q} \int p(\eta_c|\gamma) \left(\prod_{r=1}^{R_c} \sum_{z_{cr}} p(z_{cr}|\eta_c) p(w_{cr}|z_{cr}, \delta) \right) d\eta_c \tag{3.26}$$

3.3 Proposed technique

3.3.1 Problem formulation

The issue of effective cross-modal retrieval considering image and text has been addressed which involves reduction in semantic gap between text and image modality and to make a strong connection among highly related images and texts. We have a collection of images and the corresponding text in the form of paragraphs. Each image has a single text file related to it. The objective of the proposed technique is to retrieve the related texts or images given an image or text instance respectively. Let $D = (I_j, T_j, L_j)_{j=1}^{N}$ be an image-text dataset, where $I_j \in R^{d_I}$ and $T_j \in R^{d_T}$ depicts the image and text features respectively. There are total N pairs of instances. (I_j, T_j) depicts an image-text pair with same semantic label $L_j \in R^c$, where c is the number of classes of semantic concepts present in the data. As the proposed method is of unsupervised nature, so the labels are not utilized in the model training, instead they are only utilized while evaluation of performance metric for the model. Suppose $D_{train} = (I_k, T_k)_{k=1}^{N_1}$ is the training data, where $I_k \in R^{d_I}$ and $T_k \in R^{d_T}$ are respective features of image and text and N_1 represents the number of instances used in model training. Image training set is defined as $I_{train} = [I_1, I_2, ..., I_{N_1-1}, I_{N_1}] \in R^{d_I \times N_1}$ and similarly, text training set as $T_{train} = [T_1, T_2, ..., T_{N_1-1}, T_{N_1}] \in R^{d_T \times N_1}$, d_T and d_I are the dimensions of text and image features respectively, where $d_I \neq d_T$. Similar to D_{train}, $D_{test} = (I_k, T_k)_{k=1}^{N_2}$ denotes the test data, where $N_1 + N_2 = N$.

3.3.2 Hybrid SOM based cross-modal retrieval

Traditional SOM

It is also popular as *Kohonen map* after the name of its inventor *Teuvo Kohonen* who proposed it in 1982 [190]. The fundamental idea behind SOM is that the systems can be constructed to imitate the joint collaboration of the brain neurons. It is a kind of artificial neural network that follows an unsupervised machine learning approach. SOM maps the multi-dimensional input vectors (x_i) to (usually) a two-dimensional grid of nodes or neurons also known as a map. More similar inputs are linked with nodes which are closer in the grid, however, the less similar ones are associated gradually farther away [191]. The crux of traditional SOM is that each input vector is linked to that node that best matches it or the node that wins the input (also alluded as Best Matching Unit or BMU) and the subset of its spatial neighbors in the map will also get modified for better matching. One node of the map can also win over multiple inputs. SOM helps to recognize the

high-dimensional data by mapping it into a 2-D map and cluster alike data in conjunction. A traditional SOM comprises of two layers in which first incorporates nodes in input space and second constitutes the nodes in output space [192]. Figure 3.11 shows a representation of traditional SOM where an input vector from multi-dimensional space is mapped to all the neurons of the output layer of SOM but only one neuron has won over that input based upon the weight of the connection link and that neuron is also known as BMU [3]. Based upon the BMU, the weights of the neighboring neurons are also modified.

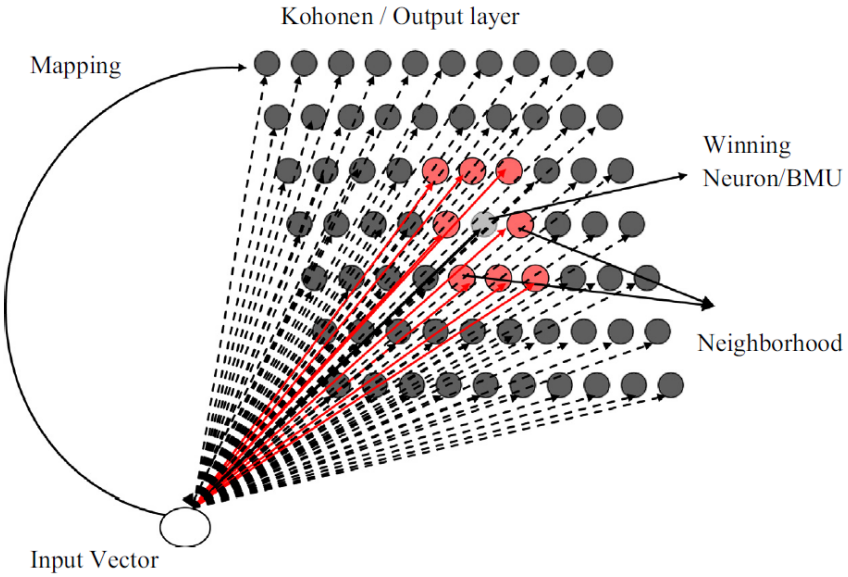


Figure 3.11: Representation of traditional SOM [3]

Table 3.1: Notations used in SOM learning algorithm

Notation	**Definition**
i	random input vector index
j	random weight vector index
x_i	input vector
w_j	weight vector of a SOM node
c	BMU index
w_c	BMU weight vector
t	index of time
$\alpha(t)$	learning weight factor
$n_{cj}(t)$	neighboring function
$N_c(t)$	neighborhood

Table 3.1 presents the variable notations and definitions which are being utilized in SOM learning algorithm. The procedure followed in traditional SOM learning is as follows [192]:

1. *Initialization*: Start with the initial values of weight vectors. Initially, each value of w_j can be picked randomly or linearly and later they will keep on adjusting with network learning process.

2. *Sampling*: Randomly select an input vector x_i from the training high-dimensional input space.

3. *Finding BMU*: Deduce the best matching unit (BMU). After comparing x_i with all the weight vectors of SOM nodes, a BMU is found lying at index c which is closest to x_i as per the Euclidean distance.

$$\|x_i - w_c\| = \min_k \|x_i - w_k\| \tag{3.27}$$

4. *Updation*: Update the BMU and its neighboring nodes. Winning node weight vector and weight vectors of its neighbors are updated as per the following equation.

$$w_j(t+1) = w_j(t) + \Delta w_j(t) \tag{3.28}$$

where $t = 0, 1, 2, ...$ depicts an index of time. The value of $\Delta w_j(t)$ is evaluated as per the following equation.

$$\Delta w_j(t) = \alpha(t) n_{cj}(t)(x_i(t) - w_j(t)) \tag{3.29}$$

where $\alpha(t) \in [0, 1]$ denotes the learning rate factor and will be decreasing monotonically while SOM learning phase. $n_{cj}(t)$ represents the neighboring function and finds the distance between nodes at indices j and c in the output layer grid. An extensively utilized neighborhood kernel is defined in terms of Gaussian function as:

$$n_{cj}(t) = exp\left(-\frac{\|r_c - r_j\|^2}{2\sigma^2(t)}\right), \tag{3.30}$$

here r_j and r_c denotes the position vectors of nodes at index j and c. The parameter $\sigma(t)$ expresses the width of the kernel which corresponds to the neighborhood $N_c(t)$ radius. $N_c(t)$ corresponds to the neighborhood set of array points around BMU (Figure 3.11). The neighborhood function $n_{cj}(t)$ value reduces while learning, from an initial value often equivalent to the

dimension of the output grid to a value equal to one.

Steps from 2 to 4 are repeated for a number of consecutive iterations during SOM learning until the weight vectors in the output layer of map represent the input patterns of high dimensional space which are closer to the map nodes, as much as possible. After initialization step, SOM learning can happen in a batch or sequential way. Both are almost similar with one difference that in sequential training, one data vector is send to the map at a time for weight adjustment rather than sending all data vectors simultaneously. After SOM training completion, each input vector is mapped to one neuron of the grid. The map size is chosen as per application. Bigger map size exposes information in detail, however, smaller map size is chosen to assure the generalization capability.

Hybrid SOM (HSOM)

In the hybrid method, two SOMs have been introduced. One SOM is dedicated to the clustering of images and another SOM is for clustering of collateral text. Each of the SOM recognizes the patterns present in the respective modalities. These two SOMs are connected to each other using a third network known as the Hebbian network which connects each node in image SOM with every node in the text SOM. Hebbian network works on the principle of Hebb's learning rule [193]. This rule is inspired by biological systems and it says that the connection between two neurons might be strengthened if they fire together. The rule states that how much the weight of a linkage between two units should be increased or decreased in proportion to the product of their activation (Eq. 3.31).

$$\Delta w_{ij} = \alpha \times x_i \times y_j \tag{3.31}$$

where w_{ij} is the weight of the link between i^{th} source unit and j^{th} destination unit, x and y represents the activities of the units. The new weight can be evaluated as (Eq. 3.32):

$$w_{ij}(n) = w_{ij}(n-1) + \Delta w_{ij} \tag{3.32}$$

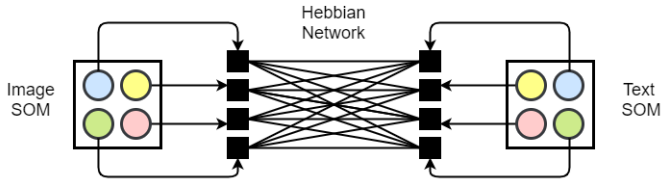


Figure 3.12: Architecture of two SOMs (image and text) connected using Hebbian network

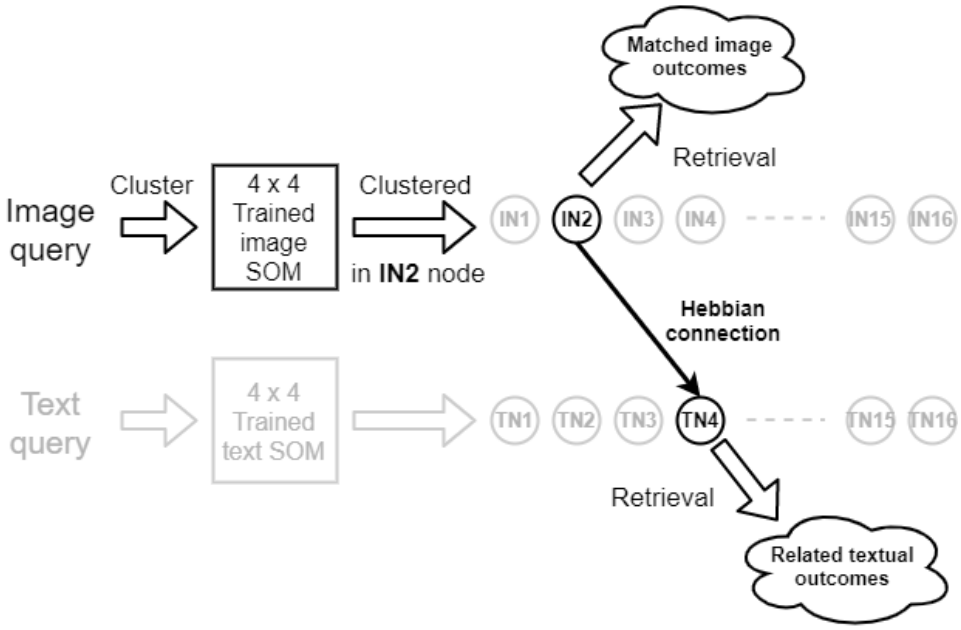


Figure 3.13: Handling of a test query. The dark portion of the figure depicts the procedure followed by an image query. Both image and text modalities are retrieved in the end.

Nodes in the two SOMs that are concurrently most active while training are associated via the Hebbian network. The purpose of utilizing the Hebbian network is to boost the connections between the two SOMs when the corresponding neurons in them activate in response to an input image and its collateral text respectively. The strength of the connection between the winning node in image SOM and between all nodes in text SOM is weighted by the activation of the connecting Hebbian node. Figure 3.12 presents the two SOMs associated with each other using the Hebbian network. If the size of image SOM is $m \times n$ and text SOM is $p \times q$, then the size of the Hebbian network would be $m \times n \times p \times q$.

For the implementation of the proposed technique, features are extracted from the available images and corresponding text as mentioned above. Two separate SOMs net_I and net_T of dimension 4×4 are trained for images and texts respectively and the node numbers are also retrieved corresponding to each instance which are saved in $classes_I$ and $classes_T$ matrices correspondingly. The trained SOM node weights $nodeWeights_I$ and $nodeWeights_T$ for both image and text SOM are fetched for further experiments. The matrices $winnersMatrix_I$ and $winnersMatrix_T$ represent the weight vector of the winner node corresponding to each image and text input instance. Afterward, Euclidean distance is calculated between each input instance vector and the corresponding winner node weight vector, and the results are saved in $whebb_I$ and $whebb_T$ which are one-dimensional matrices. Now the Hebbian network is trained using Equation 3.33 and the Heb-

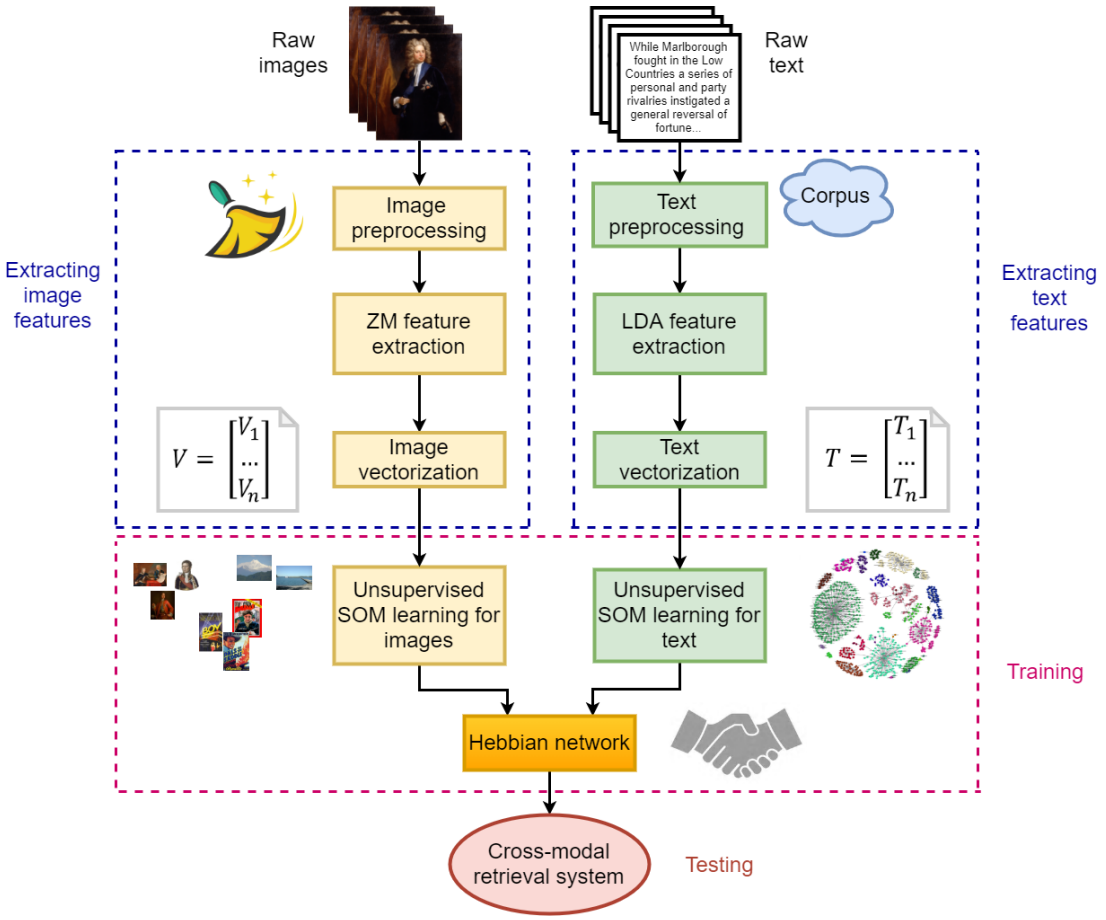


Figure 3.14: Process flow of image and text training for the proposed hybrid cross-modal retrieval system

bian link weights (depicted by *hebbLink* matrix) keep on updating for each input instance during the training process. The Hebbian network is associating each net_I node with the net_T node but the strength of the bond is determined by the link weight. The size of the matrix *hebbLink* is 16×16.

$$hebbLink(classes_I(i), classes_T(i)) += LR * whebb_I(i) * whebb_T(i) \tag{3.33}$$

where $1 \leq i \leq length(classes_I)$ and LR signifies learning rate whose value is 0.1. After the creation of the Hebbian network, two vectors $Anet_I$ and $Anet_T$ of size 16 are created such that $Anet_I$ will have the node numbers of net_T having the highest Hebbian link weight where each index of the $Anet_I$ vector represent the node number in net_I. Similarly, $Anet_T$ has the node numbers of net_I. For testing the model with new image and text instances after training, the test instance is clustered in the appropriate respective SOM node. Then the corresponding linked node in the other SOM is found and the results from both the nodes are retrieved. Thus, both image and text modality results can be retrieved using a query of any modality (image or text). The procedure of handling a test query can be

easily visualized in Figure 3.13 in which the dark portion is depicting a testing instance using an image query. The process followed in case of textual query is also similar to this one. The algorithms (1,2) present all the steps followed for the implementation of the proposed technique. Figure 3.14 demonstrates the abstract process flow of the proposed HSOM cross-modal retrieval system starting from the raw image and text data till the final trained system.

Algorithm 1 Algorithm of HSOM technique exploiting Hebbian rule for cross-modal retrieval

INPUT: D_{train} and D_{test}
OUTPUT: Trained net_I and net_T SOMs, retrieval of matched images and text corresponding to text and images in D_{test}

procedure IMAGE FEATURE EXTRACTION
  Input all images
  Resize the images to 1000×1000
  Convert each RGB image to gray scale
  Normalize the images
  Extract the Zernike moments at order 5
end procedure
procedure TEXT FEATURE EXTRACTION
  Input all XML text files
  Extract the text part from each XML file
  Decode HTML entities, remove tags, URLs and numbers from each text
  $cleanedDocuments \leftarrow tokenizedDocument(text)$ ▷ Create tokenized documents from the text
  Perform lemmatization
  Remove punctuation marks, stop words and words with $length \leq 2$
  $cleanedBag \leftarrow bagOfWords(cleanedDocuments)$ ▷ Create a bag-of-words from cleaned documents
  Find appropriate no. of topics for LDA using perplexity analysis
  Extract the LDA features.
end procedure
procedure HSOM BASED CROSS-MODAL RETRIEVAL
  Load D_{train} and D_{test}
  $dimension1 \leftarrow 4, dimension2 \leftarrow 4$ ▷ Dimensions of both image and text SOM
  $net_I \leftarrow selforgmap([dimension1 dimension2], 200)$ ▷ Configure image SOM
  $net_T \leftarrow selforgmap([dimension1 dimension2], 200)$ ▷ Configure text SOM
  $net_I \leftarrow train(net_I, I_{train}), net_T \leftarrow train(net_T, T_{train})$ ▷ Training of maps
  $classes_I \leftarrow vec2ind(net_I(I_{train})), classes_T \leftarrow vec2ind(net_T(T_{train}))$ ▷ Retrieving node number for each input instance
  for $i \leftarrow 1$ `to` $length(classes_I)$ **do** ▷ Winner node weight matrix corresponding to image input instances
    $winner_I \leftarrow classes_I(i)$
    $winnersMatrix_I(:, i) \leftarrow nodeWeights_I(winner_I, :)'$
  end for
  for $i \leftarrow 1$ `to` $length(classes_T)$ **do** ▷ Winner node weight matrix corresponding to text input instances
    $winner_T \leftarrow classes_T(i)$
    $winnersMatrix_T(:, i) \leftarrow nodeWeights_T(winner_T, :)'$
  end for
  for $i \leftarrow 1$ to $length(classes_I)$ **do** ▷ Euclindean distance calculation
    for $j \leftarrow 1$ to $imageVectorDimension$ **do**
      $whebb_I(i) \leftarrow whebb_I(i) + (winnersMatrix_I(j, i) - input_I(j, i))^2$
    end for
    for $j \leftarrow 1$ to $textVectorDimension$ **do**
      $whebb_T(i) \leftarrow whebb_T(i) + (winnersMatrix_T(j, i) - input_T(j, i))^2$
    end for
    $whebb_I(i) \leftarrow sqrt(whebb_I(i))$
    $whebb_T(i) \leftarrow sqrt(whebb_T(i))$
  end for
  for $i \leftarrow 1$ to $length(classes_I)$ **do**
    Train the Hebbian network using Equation 3.33
  end for
  Follow Algorithm 2 for creation of $Anet_I$ and $Anet_T$
  Cluster $I_k \in net_I$ and $T_k \in net_T$ where $(I_k, T_k) \in D_{test}$ and $k \in [1, N_2]$
  Refer $Anet_I$ and $Anet_T$ to find the corresponding Hebbian link node
  Retrieve results from the *found* node
end procedure

Algorithm 2 Algorithm for creation of $Anet_I$ and $Anet_T$ vectors

INPUT: Trained Hebbian Network
OUTPUT: Two 1-D vectors $Anet_I$ and $Anet_T$ of size 16 each

1: **procedure** CREATION OF $Anet_I$
2: $net_ISize, net_TSize \leftarrow dimension1 \times dimension2$ ▷ Size of image and text net (net_I, net_T) in Hebbian network
3: **for** $i \leftarrow 1$ to net_ISize **do**
4: $maxtemp \leftarrow 0$, $maxindex \leftarrow -1$ ▷ Initializing the temporary variables
5: **for** $j \leftarrow 1$ to net_TSize **do**
6: **if** $hebbLink(i, j) > maxtemp$ **then** ▷ Checking for the maximum hebbLink weight
7: $maxtemp = hebbLink(i, j)$
8: $maxindex = j$
9: **end if**
10: **end for**
11: $Anet_I(i) = maxindex$
12: **end for**
13: **end procedure**
14: **procedure** CREATION OF $Anet_T$
15: **for** $i \leftarrow 1$ to net_TSize **do**
16: $maxtemp \leftarrow 0$, $maxindex \leftarrow -1$ ▷ Initializing the temporary variables
17: **for** $j \leftarrow 1$ to net_ISize **do**
18: **if** $hebbLink(j, i) > maxtemp$ **then** ▷ Checking for the maximum hebbLink weight
19: $maxtemp = hebbLink(j, i)$
20: $maxindex = j$
21: **end if**
22: **end for**
23: $Anet_T(i) = maxindex$
24: **end for**
25: **end procedure**

3.4 Conclusion

This chapter introduced new ways of intelligently training neural computing systems and querying them using images or text to retrieve matched texts or images respectively. The visual features extracted from images are Zernike moments that have almost no redundancy. LDA features are considered as the linguistic features for the text. Two unsupervised traditional self-organizing feature maps are trained simultaneously but separately for images and collateral text respectively. A Hebbian link is set up between the most active nodes in the two SOMs. This is the basis of our claim that we use multi-modal features for training neural networks and also establish cross-modal links between the two maps using a Hebbian network while the training process. In reality, getting a labeled data is quite difficult, so the proposed framework will work effectively in that case as it is of unsupervised nature and thus does not require any data labeling.

Chapter 4
Image Clustering using WT, ZM, and SOM

4.1 Overview

The outline of the chapter is mentioned in the following points:

1. A novel algorithm for medical image clustering has been proposed which is based on unsupervised neural classifier systems.

2. The characteristic visual features are obtained from the images using Wavelet Transforms (WT), Zernike Moments (ZM), and Kohonen self-organizing feature map algorithm has been applied for clustering.

3. The proposed image clustering approach has been applied to the real capsule endoscopy images obtained from a known gastroenterologist and data distribution has been carefully studied using PCA and LDA plots to motivate the application of advanced machine learning techniques.

4. Performance analysis of the proof-of-concept model has been compared with both traditional and contemporary methods to support the belief.

4.2 Proposed scheme

Automatic image analysis and segmentation is a skilled task carried out by experienced professionals. Features in an image are used to decompose and analyze the underlying anatomy by defining a mechanical and systematic procedure. Given the explosive growth of visual information, partly due to the expansion of the Web and partly due to the introduction of sophisticated and inexpensive image capture systems, there is an urgent need to develop programs that can learn to segment and annotate. Automatic segmentation and annotation systems are among the critical areas of research and development for the next decade and beyond, and machine learning will be a vital technology in developing such systems [194, 195]. Self-organizing maps (SOM) incorporated with extended fuzzy c-means clustering have been a popular method for image segmentation as studied in [196]. It has used a discrete wavelet transform for image description for edges and lines involved in contrast variation.

The objective of the proposed study in this chapter is to analyze, segment, and cluster the endoscopy images such that the trained system can be helpful for gastroenterologists in problem diagnosis of the gastrointestinal tract. The motivation behind the current problem selection is its complexity in terms of image feature distribution. An example of in-vivo gastral images has been shown in Figure 4.1 in which an image has been analyzed using two segmentation algorithms: (a) *Region Growing* (It has been applied in [197] to segment 2D microscopy digital images of freshwater green microalgae. In this approach, the image is segmented into multiple disjoint regions (sub-regions), and then they are merged with their nearest neighboring seeded region (to grow regions) that satisfies a predefined homogeneity criterion.); and (b) *2D Otsu algorithm* [198] (which employs the gray level information of each pixel and its spatial correlation information within the neighborhood). The algorithm has failed to capture the region of interest in both the cases, which is bleeding and not the dark spot.

It can be observed that it is pretty challenging to accurately segment blood due to the obscure nature of the color distribution and irregular region boundary. The red and green boundaries have captured the wrong dark region instead of the red spot ROI (region of interest). Moreover, the underlying images are dynamic, involving continuous movements of the camera in the drifting capsule, body organs, insufficient light conditions to capture texture at the region of interest, and varying luminance and noise due to food particles and body fluid. In addition, complementary metal-oxide semiconductor (CMOS) image sensors involve noise, high compression ratio, and low resolution of 256×256. If a segmentation method can enhance the classification accuracy in this confounding case, then inherently, it would also contribute to other applications of image processing. This is the reason for the underlying case study about image segmentation for gastral images. Challenges involved in image retrieval have been discussed in Table 4.1.

We have used wavelet resolution which helps to remove noise and makes images scale invariant. Zernike moments have been used for image vectorization and self-organizing maps based on unsupervised learning is used to cluster images for sick and healthy classes.

4.2.1 Background

This section summarizes the miscellaneous works by various researchers related to the proposed work. Muhammad et al. studied a comprehensive survey of computer vision techniques for wireless capsule endoscopy (WCE) [199]. Information regarding various publicly available datasets of WCE has also been provided along with challenges and future scope. A survey has been presented in [200] for includ-

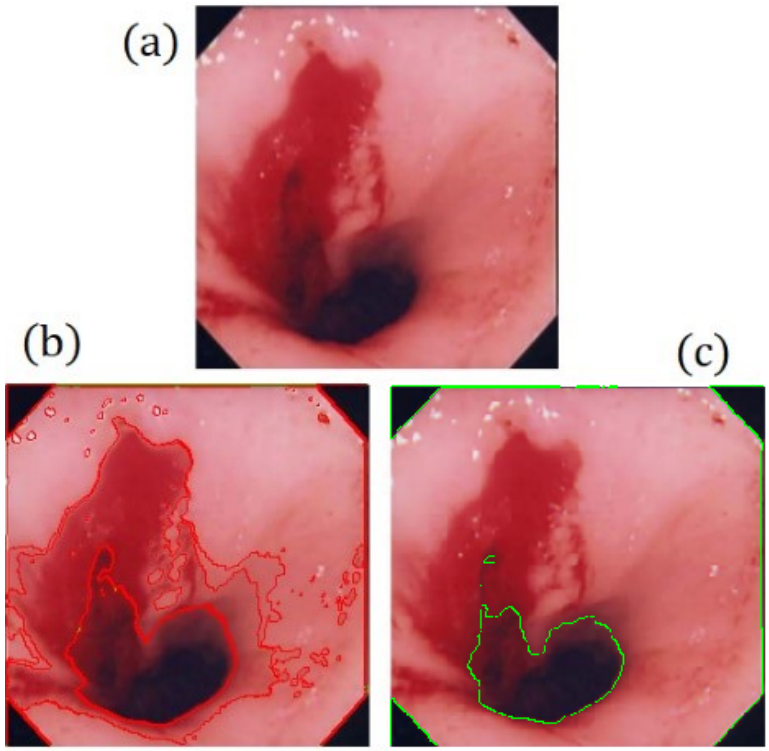


Figure 4.1: (a) Original gastral image, (b) Region growing segmentation [197], and (c) 2D Otsu segmentation algorithm [198]. The red and green boundaries have captured the wrong (dark) spot instead of red region of interest (ROI) showing that problem is complex for image segmentation.

Table 4.1: Summary of challenges of image representation and learning

Challenge	**Elaboration**
Image invariance	Yields same image, when rotated, scaled or moved.
Noise	The 'lens' of the camera is never perfect; surrounding environment may contribute to the noise, noise could be Gaussian or distributed differently.
Representation	In terms of the optical properties of the (individual) pixels of an image – mean intensity, x-tilt, y-tilt, focus astigmatism @ 0 degree & focus astigmatism @ 45 degrees, coma & x-tilt, coma & y-tilt, spherical & focus.
Learning	For recognizing the contents of a new image having "see" similar images before.

ing deep learning to automate the process of WCE examination. Deep learning applications for WCE such as detecting polyps, bleeding, ulcers, hookworm, and celiac disease are discussed. A computer-aided diagnosis technique has been proposed in [201] for identifying and categorizing the abnormalities in vision-centered endoscopy detection. A novel deep sparse SVM feature selection model with group

sparsity has also been incorporated, which assigns an appropriate weight to the feature dimensions and also removes the inadequate features from the feature pool. Radhika et al. have utilized Zernike moments (ZM) to authenticate online signatures, and ZM represents the shape of the acceleration plot [202].

A novel recurrent framework has been proposed in [203] for joint unsupervised learning of deep representations and image clusters. The sequential tasks in the clustering algorithm are expressed as steps in the recurrent process, stacked on top of convolutional neural network (CNN) representations output. The research is inspired by the fact that good representation benefits image clustering, and clustering output gives supervisory indications to representation learning. Zhu et al. have proposed a Nonlinear Subspace Clustering (NSC) technique for image clustering that exposes the multi-cluster nonlinear structure of data instances using a nonlinear neural network [204]. The technique introduced in [205] quantifies the clusterability of a dataset and is based on the probability density of a measure (S) of clusterability (in 1D) of projection of data onto a random line. After comparing the clusterability of image datasets with synthetically created clusters, it has been inferred that the structures we discover in image datasets do not fit the notion of clusters in the traditional sense. Moreover, the authors introduced a fast approach to hierarchically clustering high-dimensional data. Chang et al. have proposed a Deep Adaptive Clustering (DAC) approach to represent the clustering problem as a binary pairwise classification framework for identifying whether pairs of images belong to the same cluster [206]. The cosine distance metric has been utilized for calculating the similarities between label features of images produced by a deep convolutional network.

A novel technique, Robust learning for Unsupervised Clustering (RUC), has been introduced in [207] that is motivated by robust learning and overcomes the issues of faulty predictions and overconfident results in the case of unsupervised image clustering. This approach utilizes the pseudo-labels of existing image clustering models as noisy data that may comprise misclassified instances. Ren et al. have proposed a two-stage deep density-based image clustering (DDC) framework to address the issue of selecting an appropriate number of clusters in advance [208]. A pseudo-supervised joint approach has been proposed in [209] for image clustering, named Discriminative Pseudo Supervision Clustering (DPSC). Authors have resolved two significant issues in image clustering problems: appropriate image representation and lack of supervision. The main idea is to determine and use the pseudo supervision information for providing supervisory guidance for discriminative representation learning.

An improved version of ZM has been introduced in [210], which has been utilized for face recognition. In addition to the basic orthogonal and intrinsic

characteristics, this version is also invariant to noise, illumination, translation, in-plane rotation, and scaling. A hybrid similarity measure has also been proposed in this by integrating Jaccard similarity with L1 distance. Fractional-order Zernike moments, an improved version of ZM, have been presented in [175] for analyzing the grape leaf images. Multi support vector machine classifier is utilized to classify grape leaf diseases. Daubechies complex wavelet transform (DCxWT) and ZM have been used in combination for image representation in [211]. The multi-class support vector machine is used for object classification. To denoise image sequences using nonlocal means extended by ZMs, is proposed by [212]. It is found to be faster due to a reduction in weight computations, and block matching has been discounted. Similarity distance is found using photometric distance in consecutive images.

A local ZM based spatio-temporal feature is proposed in [213] in the spatial domain exploiting motion change frequency for recognizing facial expressions. In Yang et al. performed a study of modified principal component analysis to extract image features from the ORL face database and has been named image projection PCA (IMPCA) [214]. Sparse coded features are introduced for identifying bleeding in wireless capsule endoscopy images in [215]. These features are obtained after computing Scale-Invariant Feature Transform (SIFT) and uniform Local Binary Pattern features for WCE images. SVM is utilized for classifying the images. In Gupta et al. have proposed an automated system for detecting focal electroencephalogram (EEG) signals by using differencing and flexible analytic wavelet transform (FAWT) techniques [216]. K-nearest neighbor and least squares support vector machine are applied as classifiers for automatic diagnosis.

Table 4.2: Summary of literature survey: pre-processing and noise removal, image representation, and learning

Sr	Method	Purpose	Outcome	Study
Pre-processing and noise removal				
1	DCxWT, ZM and multiclass SVM	Object classification	Better precision and accuracy values	Khare 2021 [211]
2	Nonlocal means extended by ZMs	Faster computation	Denoising and faster computation	Singh 2017 [212]
3	Differencing and FAWT	Automatic detection of focal EEG signals	94.41% accuracy	Gupta 2017 [216]
4	Riesz wavelets	image retrieval for a hemangioma, liver lesions	NDCG score = 0.92, AUC = 0.77	Kurtz 2014 [217]
Image representation				
5	Zernike moments	Online signature authentication	4% of False Rejection Rate, 2% of False Acceptance Rate	Radhika 2011 [202]

6	Local modified Zernike moment per unit mass	Face recognition	Higher recognition accuracies on two datasets	Kar 2020 [210]
7	Deep sparse SVM	Computer aided endoscopy diagnosis	New endoscopy dataset, Computation reduction and improved robustness	Cong 2015 [201]
8	Image principal component analysis	To analyse IMPCA is better than PCA, FDA	Better accuracy and reduced time	Yang 2002 [214]
Learning				
9	Local ZM, SVM	Facial expression recognition	Improved recognition rate	Fan 2017 [213]
10	Survey of computer vision methods for WCE	Determining major challenges of WCE and future scope	Comparative analysis	Muhammad 2020 [199]
11	Survey of deep learning for WCE	Systematic review and meta-analysis of deep learning methods for WCE	Comparative analysis	Soffer 2020 [200]
12	Sparse coded features, SVM	Detect bleeding in WCE	accuracy = 98.18%	Patel 2021 [215]
13	DWT, Invariant moments, ANN, KNN	veterinary field, spermatozoa healthy or sick	accuracy = 95%	Alegre 2012 [218]
14	Local spatial features, Rayleigh PDF model	Automatic bleeding detection in WCE images	Improved performance with less complexity	Kundu 2019 [219]
15	Stacked sparse autoencoder with image manifold constraint	polyp recognition	Overall accuracy = 98%	Yuan 2017 [220]

Alegre et al. predicted an automatic quality assessment of sperm quality (damaged or intact) using ANN and KNN [218]. Co-occurrence matrix and discrete wavelet transforms have been calculated from the underlying images for texture features and have been found to outperform moment-based descriptors in the study. A probability density function (PDF) based approach has been proposed in [219] for automatic detection of bleeding in WCE images. After determining the pixels of interest, local spatial features are extracted from the images by employing a linear separation scheme.

Kurtz et al. studied an image retrieval system based upon semantic features [217]. It uses ontological terms to define the image using multi-scale Reisz wavelets to analyze their annotation similarity. Liver lesions in CT images have been experimented with to validate the proof-of-concept. Normalized discounted cumulative gain (NDCG) score and AUC have been calculated and compared for the real-time

decision-making capabilities of the model. For the robust representation of WCE images, the study given in [220] provides the assistance and discriminated definition for polyp images using a deep learning technique utilizing sparse auto-encoder. It uses a nearest neighbor graph to define inherent image manifold characteristics. A summary of the motivational literature review has been given in Table 4.2.

4.3 Image feature vectors

Image features involve color, texture, and shape metrics based upon the contrast-related discontinuities in the image. For this study, Wavelet Transforms [221] and Zernike moments [222] together have been used due to their efficiency and power to capture the inherent characteristics.

4.3.1 Wavelet transforms (WT)

These mathematical functions divide a signal (image) into different frequency components. The goal is to study each component with a resolution with a matching scale. WT is better than Fourier Transforms (FT) or Short-Time Fourier Transform (STFT), which cannot analyze both frequency and time components [223]. Wavelet transform is composed of wavelet function $w(.)$, defined in finite time and normalized. The formula for WT is:

$$W_{f(\mu,\sigma)} = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{\sigma}} w\left(\frac{x-\mu}{\sigma}\right) dx \tag{4.1}$$

where (μ, σ) are translation and scaling parameters, respectively. To see lower frequency components of the signal, increase the value of σ for instance. Some prominent mother wavelets have been shown in Figure 4.2. In our study, Daubechies 4 wavelet has been used (details in [224]). Spatial information comprises the image pixel positions (x, y) that act as the time axis and changes in pixel intensity $f(x, y)$ that serve as the frequency axis. Thus, edges have a higher frequency as compared to smooth areas. For Discrete WT (DWT), an image is decomposed into four components: approximation, horizontal, vertical, and diagonal. As shown in Figure 4.3, the image is decomposed into one level in (4.3a) and two levels in (4.3b) with high/low pass filters and down sampling the signal (image).

In our study, the image has been decomposed on three levels using WT, as shown in Figure 4.4. It explains about horizontal, vertical and diagonal edges being detected in the original image. Ten components have been calculated as $\{(H_i, V_i, D_i, A_i) \mid i = 1, 2, 3 \text{ for } H, V, D \text{ and } i = 4 \text{ for } A\}$. In expanded form, we get $H_1, V_1, D_1, H_2, V_2, D_2, H_3, V_3, D_3$, and A_3. Then for these 10 components,

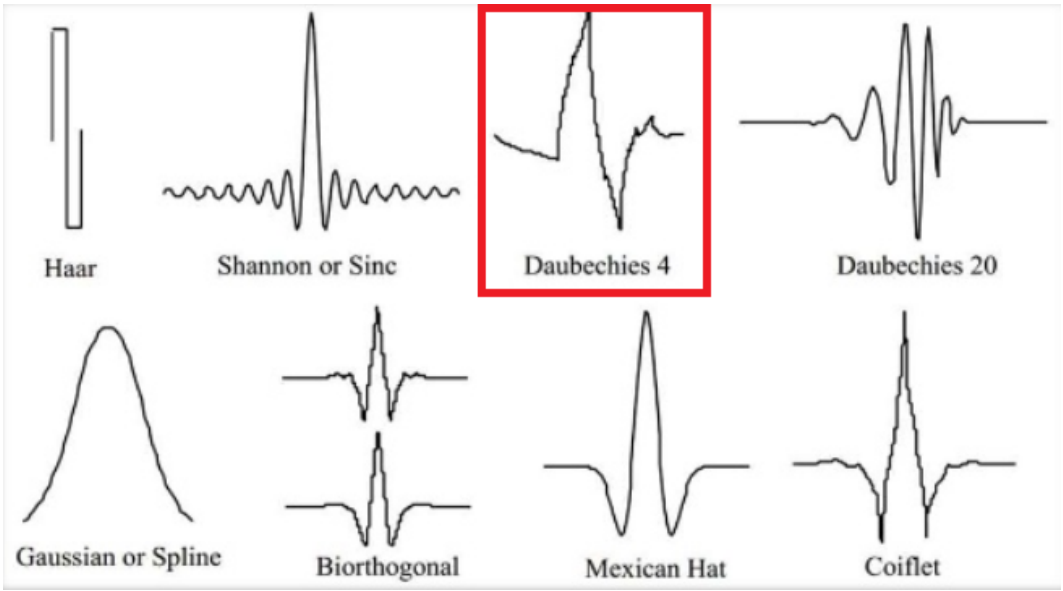


Figure 4.2: A few popular mother wavelet functions $w(.)$. Daubechies 4 wavelet have been utilized in the experimentation.

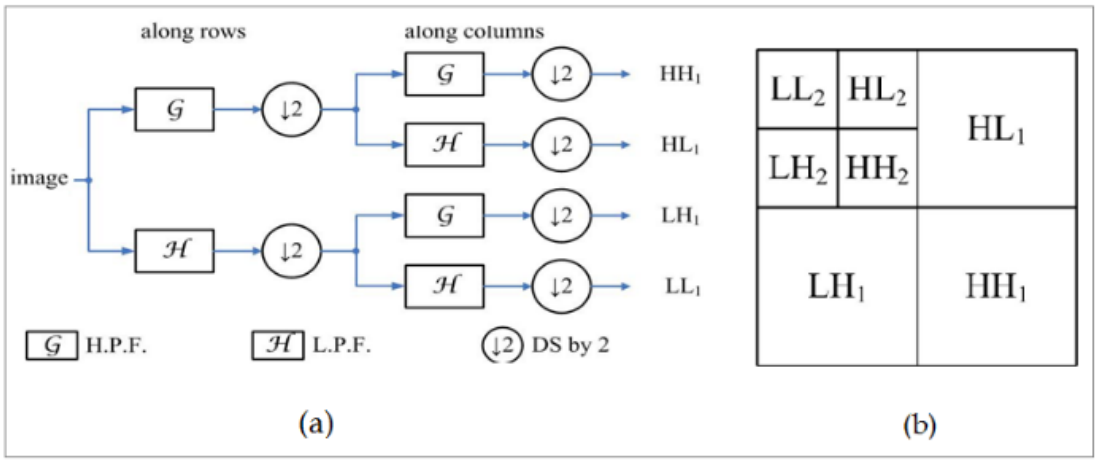


Figure 4.3: Image decomposition using DWT. HPF – High Pass Filter, LPF – Low Pass Filter, DS – Down Sampling, (a) Image decomposition into HH1, HL1, LH1, LL1 into two dimensions; (b) Next level 2 analysis (2-L 2-D) with DWT sub-band [225].

12 Zernike moments have been calculated for $n = m = 5$ which are listed as $Z_{00}, Z_{11}, Z_{20}, Z_{22}, Z_{31}, Z_{33}, Z_{40}, Z_{42}, Z_{44}, Z_{51}, Z_{53}$, and Z_{55} or in the set notation $\{Z_{ij} \mid i \geq j \text{ and } i - |j| \text{ is even}\}$. After compiling all that information from LUV channels, the image feature vector has $12 \times 3 = 36$ dimensions. For example, Figure 4.5 shows the results from a sample picture's approximation, horizontal, vertical, and diagonal edge detection decompositions.

Wavelet Transformation (WT) is quite useful for noise removal, image compression [221], and zooming capabilities for local characteristics of an image. It is also an efficient technique for texture characterization while preserving local and

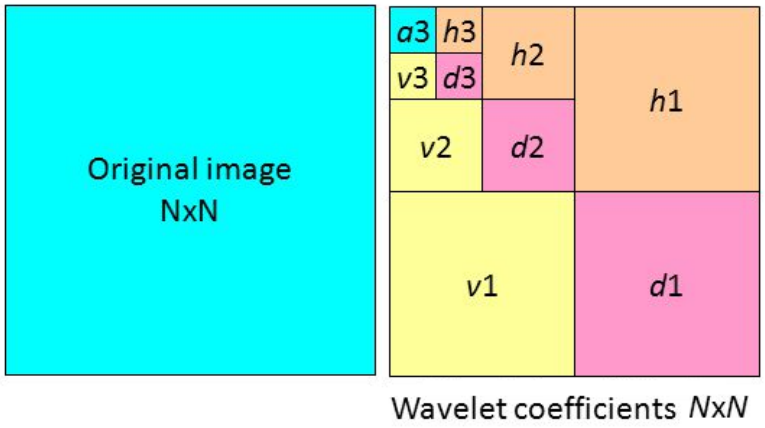


Figure 4.4: Daubechies wavelet transformations are used in the experiments. a, v, h, d stands for approximation, vertical, horizontal, and diagonal details. Diagonal (low/low), horizontal (high/low), vertical (low/high), approximation (high/high).

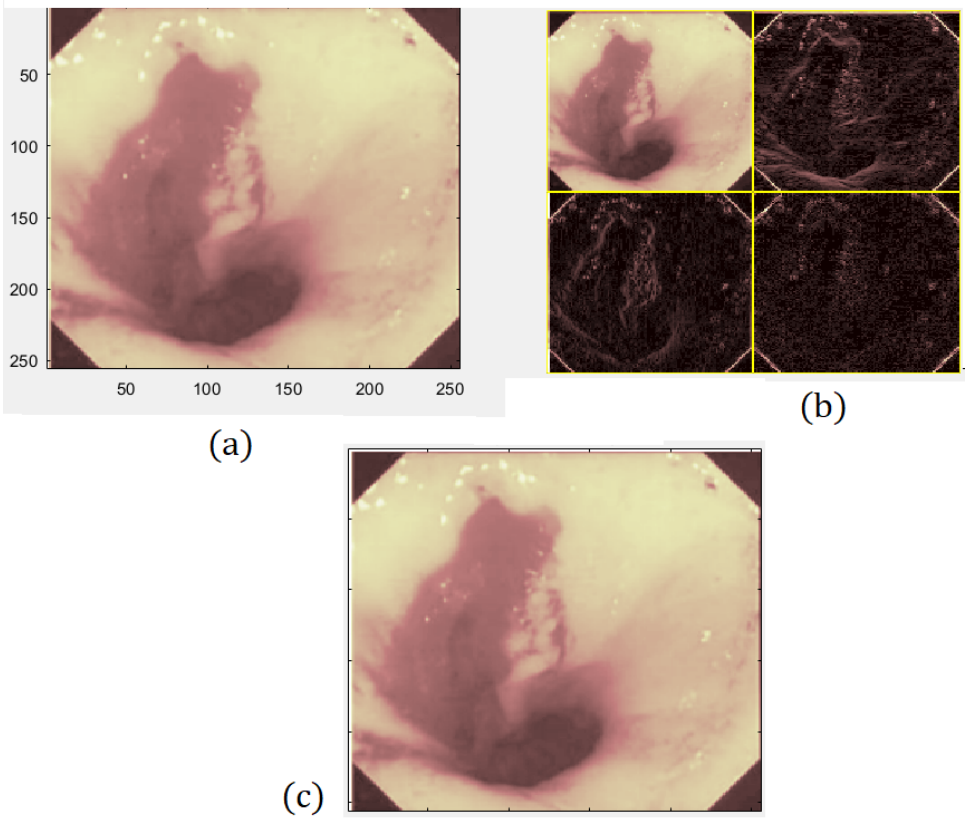


Figure 4.5: The four decompositions explained with example: approximation, horizontal, vertical, and diagonal details to detect the corresponding edges. Figure (a) is original image, (b) is the view of four decompositions, and (c) is denoised image.

global spatial/spectral information. For instance, the noise removal feature of WT is shown in Figure 4.5 with four decompositions levels, and image denoising has been illustrated in Figure 4.6 for an image with a considerable amount of Gaussian

noise.

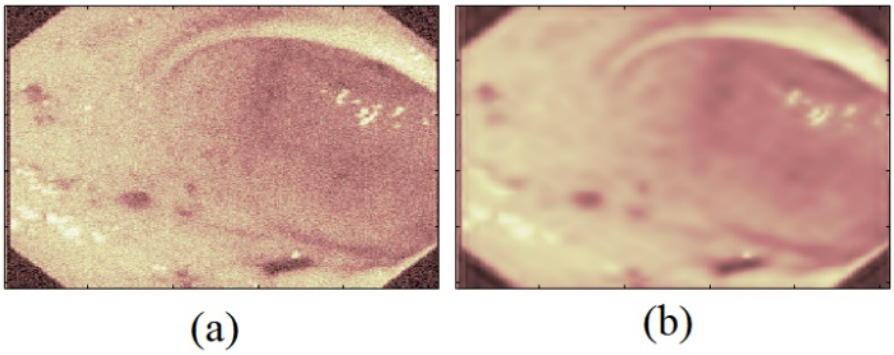


Figure 4.6: (a) Noisy image and (b) denoised image with Daubechies wavelets (DB-4).

4.3.2 Zernike moments (ZM)

Image moments are the weighted average of the intensity values of the image pixel (or a similar image function) to get the scalar quantities for image interpretation. Moments of different order yield varying information about the image, such as area, center of mass, and orientation. Zernike Moments (ZM) [169] of an image are similar to Discrete Cosine Transform (DCT) coefficients in their derivation and properties. ZM are projections of an image function along the real and imaginary axes (x-axis and y-axis), which are convolved by an orthogonal function. They represent an image in various frequency components which are referred to as the orders (along the radial) and repetitions (along the angular direction). Thus, Z_{00} represents the average intensity, Z_{11} represents the first-order moment, Z_{20} is similar to variance, and so on. Zernike polynomials are orthogonal functions that generate an orthogonal set over the unit circle in a complex plane. The center of the image stays the same as the center of the circle. Hence, a square image can be mapped inside or outside an image [184]. In the case of inner mapping, the pixels which fall outside the unit disc must be discarded. So, to avoid the information loss from the edges, we have utilized outer mapping for our experimentation. The detailed description and formulae for ZM are mentioned in chapter (3).

4.4 Self organizing map

Our method involves definitions for creating a set that associates the most active neuron for the set of the output layer of SOM, with a set of input vectors presented to the input layer of SOM as defined in equations (4.4 and 4.5). It applies to a

single SOM or can be extended as the collateral SOM for hybrid SOMs. Follow Algorithm 3 for creating single modal information systems for image clustering.

Algorithm 3 Algorithm for retrieving information from a single SOM

1: Identify the best match node $\vec{w}_k$.
2: Form a totally ordered set of the n nodes in the SOM, such that:

$$(W, \leq) = \left\{ \begin{array}{c} \vec{w}_k, k = 1..n \mid \vec{w}_i \leq \vec{w}_j \Leftrightarrow \\ \|\vec{x}_k - \vec{w}_i\| \leq \|\vec{x}_k - \vec{w}_j\| \end{array} \right\} \tag{4.2}$$

where $\vec{w}_i, \vec{w}_j \in W, 1 \leq i, j \leq n$ and $i \neq j$
3: Retrieve a totally ordered set R, of all p pre-stored items used in training, in response to the input vector $\vec{x}$

$$R_{single} = \{\vec{x}_l, l = 1..p \mid \exists\, \vec{w}_k \in W : (\vec{x}_l, \vec{w}_k) \in P_{single}\} \tag{4.3}$$

A Self-organizing Map (SOM) [191], also called a Kohonen Map, associates a multidimensional input space, comprising a set of feature vectors, onto a 2-dimensional surface (output map). The end of training leads to an association between an input vector $\vec{x}$ and a specific output node that 'wins' the input, known as the Best Matching Unit (BMU) for that input vector. If $\vec{w}$ represents the weight vector of an output node, then BMU for input vector $\vec{x}$ can be calculated as:

$$\|\vec{x} - \vec{w}_m\| = min\{\|\vec{x} - \vec{w}_m\|\} \tag{4.4}$$

where m depicts the index of SOM output node which is a BMU. One node may 'win' over more than one input forming a set. Let P_{single} be the pair set of q input vectors and the corresponding winning node is $\vec{w}_m$, then P_{single} is defined as:

$$P_{single} = \left\{ \begin{array}{c} (\vec{x}_k, \vec{w}_m), k = 1..q \\ \|\vec{x}_k - \vec{w}_m\| = \min\limits_{i=1}^{n}\{\|\vec{x}_k - \vec{w}_i\|\} \end{array} \right\} \tag{4.5}$$

Information retrieval from a SOM involves the presentation to the trained SOM of a set W. The mapping of the input vector from higher dimensional nodes in the output layer forming a space to the winning node in 2-D neuron space has shown in Figure 4.7. The length of input vector X_i and neuron weight vector W_i must be the same. The retrieving information from a SOM has been depicted in the Algorithm 3. The following section explains image vector creation using Zernike Moments and Wavelet transformation for denoising.

During the initial stages of the SOM training, the weight vectors are initialized with random weights and then, together with the input vectors, are normalized.

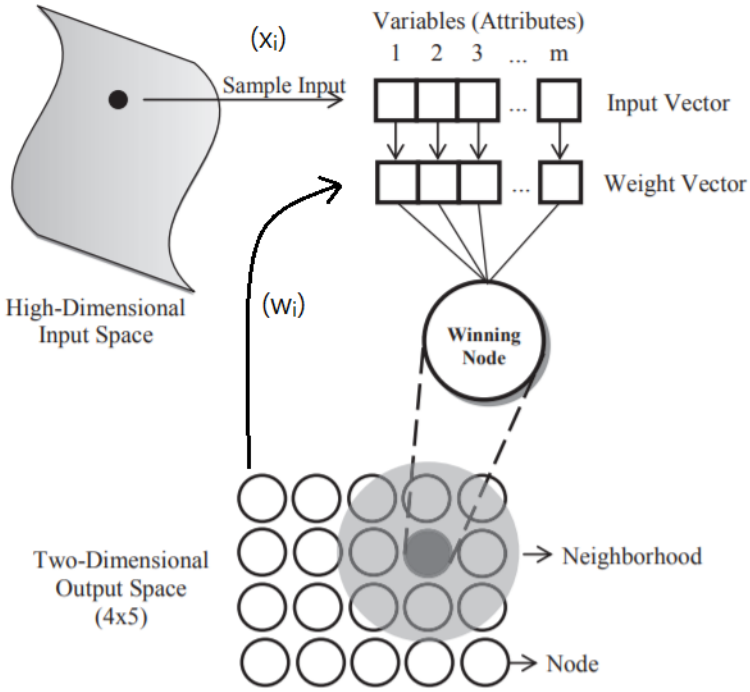


Figure 4.7: Overview of single self-organizing map (SOM) model. X_i are input vectors with same length as weight vectors W_i. Each X_i is connected to every (winning) neuron.

The learning and neighborhood rates are reduced exponentially during training following established practice in the SOM literature. Our testing regimen relies on the notion of best matching unit(s): the node(s) in the output layer that responds with the highest activation value to a given input vector. Note that if one or more neurons can be activated in response to the input vector, then the activated neurons are ordered according to their activation levels (algo 3). If the category of the input vector matches the most activated neuron in the output layer, then we have a best-matching unit (BMU). If there are multiple activated nodes for a specific input then we are considering the two highly activated nodes only.

A matching matrix was created to analyze how an input vector may activate neurons that were trained to respond to one or more categories of keywords or images. If the winner or BMU in the output layer has the same category as the stimulus, and the stimulus did not excite any other neurons, then the match will be perfect. However, if a given stimulus activates neurons of various other categories, the match will be minimal. We define accuracy as the number of correctly clustered items (based upon the majority of similar items in the cluster as the test instance) divided by the total number of items in the category.

4.5 Proposed clustering technique

The proposed research aims to effectively cluster the in-vivo gastrointestinal images based upon their similarity by carefully considering the image semantics. Let I be the training set of images that is an input to the proposed algorithm. The expected output is the trained self-organizing map and the image cluster sets (C_i) constructed as per the image similarity. The first step is to denoise the images using Daubechies wavelets with four decomposition levels: approximation, horizontal, vertical, and diagonal. The next step is the conversion of RGB to LUV channels. Wavelet transforms implementation details are given in *Section 4.3.1.* Subsequently, 12 Zernike moments are calculated for each of the L, U, and V channel with $n = m = 5$, creating a total of $12 \times 3 = 36$ image vector dimensions. The ZM calculation steps and equations are mentioned in detail in *Section 4.3.2.* In the end, 4×4 SOM is trained using image vectors and constructs the image clusters. Algorithm 4 shows the steps for segmentation and clustering of the images using SOM.

Algorithm 4 SOM image clustering algorithm

INPUT: Training set of images I

OUTPUT: Image cluster sets (C_i where i = number of SOM clusters and $C_i \subseteq I$)

1: **procedure** SOM TRAINING FOR ENDOSCOPY IMAGES
2: Image denoising - Daubechies wavelets ($DB-4$) using four decomposition levels (a, v, h, d)
3: RGB to LUV transformation
4: Calculate ZM for each of L, U, and V channels with n=m=5
   (i) Calculate radial polynomial $R_{st}(r)$ using eq. 3.6
   (ii) $V_{st}(x, y) = R_{st}(r)e^{it\mu}$
   (iii) Z_{stx} and Z_{sty} are real and imaginary values of Z_{st}
   (iv) $Z_{st} = \sqrt{Z_{stx}^2 + Z_{sty}^2}$
   (v) Calculate 12 Z_{st} for each L, U, and V channel, so total 36 elements in image vector
5: Train SOM with 4×4 grid size using the obtained image vectors
6: Required image clusters (C_i) are obtained after SOM training
7: **end procedure**

Figure 4.8 is the pictorial representation of all the steps involved in implementing the proposed approach. Initially, we have a set of 300 raw endoscopy images. The images are pre-processed and the region of interest is identified. Afterward, the denoising of the images is performed using wavelet transforms and the RGB images are converted to LUV format. Subsequently, ZM features are extracted

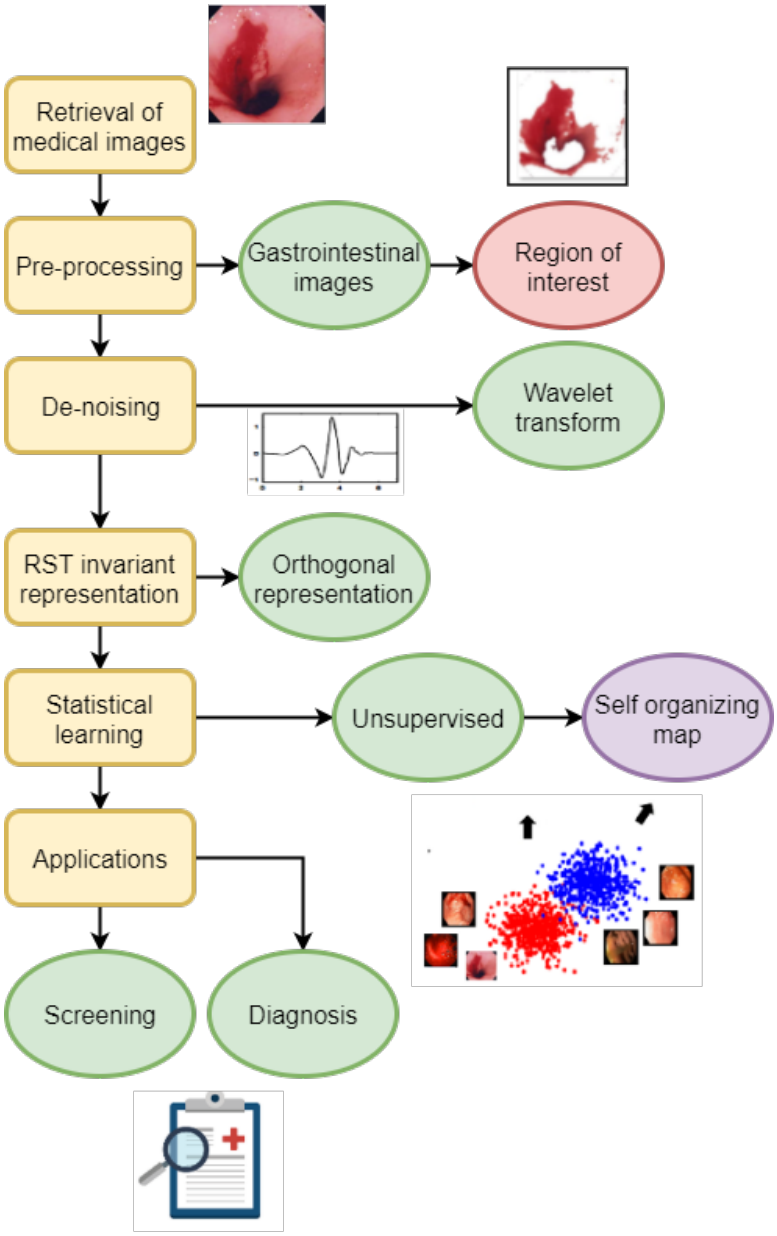


Figure 4.8: Pipeline diagram for the proposed methodology

from the images, which are rotation, scaling, and translation invariant. The unsupervised self-organizing map is trained using the extracted image features, and the image clusters are formed based on the similarity. Now the trained system can be utilized by gastroenterologists for screening and diagnosis purposes for endoscopy images.

4.6 Conclusion

This chapter introduced new ways of intelligently segmenting and analyzing image collections by training neural computing systems with images having obscure

color and texture contrasts. The characteristic visual features of the image collection are derived from Wavelet Transforms and Zernike Moments. The images are categorized using an unsupervised clustering algorithm – Kohonen self-organizing feature maps. The proposed system can classify sick and healthy in-vivo images effectively without the labeled data, which is hard to get in reality, specifically medical data. It is often expensive to manually label the data by an expert in the related field. The system is beneficial in clustering vague color distribution, asymmetrical region boundary, and noisy image data. It is rotation, scaling, and translation invariant due to the use of ZM for image representation.

Chapter 5
Cross-modal Retrieval using Oja Rule

5.1 Overview

The following points give the chapter overview:

1. The proposed approach associates the image and text modalities for cross-modal retrieval process.

2. The training of two SOMs is performed independently using deep image features and TFIDF text features.

3. The highly correlated image and text SOM neurons are integrated together using the Oja rule.

5.2 Image feature extraction

A classification experiment has been performed with several pre-trained deep convolution neural networks for selecting the appropriate network for deep visual feature extraction. After feature extraction, images are classified into respective classes using a support vector machine (SVM) classifier. Out of all the deep model features, SVM gave the highest accuracy using VGG16 features, so it has been utilized in the proposed approach for image feature extraction. Table 5.1 shows the models chosen for experimentation along with the classification accuracy. VGG16 and the corresponding accuracy value are represented as bold in the table because of the highest value. VGG16 convolution neural network has been used for classification or visual feature extraction in many applications recently and has also been found to be effective [226, 227, 228].

5.2.1 VGG16

VGG16 network is developed by the *Visual Geometry Group* of the University of Oxford, and it is the winner of the 2014 ILSVRC object identification algorithm [229]. It is a deep convolutional neural network pre-trained on the ImageNet dataset comprising more than a million images. It is capable of categorizing the images into 1000 object classes. Hence, the network has learned rich feature

Table 5.1: Different deep learning models chosen for experimentation along with classification accuracy (using SVM) to select the best accuracy model for image feature extraction.

Model	**Accuracy**
alexnet	0.7667
vgg16	**0.8333**
googlenet	0.6667
squeezenet	0.6667
inceptionv3	0.6333
densenet201	0.8
mobilenetv2	0.6667
resnet18	0.6333
resnet50	0.7667
resnet101	0.7
inceptionresnetv2	0.6

representations for a variety of images. VGG16 takes the fixed input of 224×224 RGB image. Its network architecture consists of a total of 41 stacked layers. There are 16 layers with learnable weights: 13 convolutional layers and 3 fully connected layers. A kernel of 3×3 dimension is utilized for the convolution operation along with W and b (as learnable attributes), which are passed over the pixels x of an image, and it gives y as the output. The following equation simply represents the convolution task by the function:

$$y = f(Wx + b) \tag{5.1}$$

The convolution layers extract patterns to distinguish among different classes. Simple features learned by initial convolution layers are combined to create complex features in the later convolution layers. Rectified Linear Unit (ReLU) activation layer is typically placed after each convolution layer to introduce uncertainty. The maxpooling layer performs downsampling to reduce the activation map size. A classifier exists at the end of this stack of convolution layers. There are two fully connected layers of 4096 neurons and one fully connected layer of 1000 neurons after these. The output from this layer goes into the softmax layer, which gives a probability score for each category. Then the classification layer (last layer) assigns it to a category as per the cross-entropy function. In the proposed study, VGG16 has only been utilized for feature extraction, so both the softmax and classification

layer are absent. Image features are retrieved from the last fully connected layer of 1000 neurons, creating a feature vector of 1000 dimension corresponding to each image.

5.3 Text feature extraction

For text vector creation, TFIDF features have been extracted which are widely used in document representation. These features identify the importance of each word to a document in a collection and give less weightage to words that appear more often. Given that each image can be associated with a set of, say, up to n keywords, if the image collection comprises images in different categories, then some of the collateral keywords, relating to the individual categories, will be shared across a number of images – so we may have n unique keywords. Some of the keywords are used very frequently, some less so, and some very rarely.

Frequency-based metrics are typically used to construct vectors for sets of text documents: the vectors comprise information about the presence or absence of *significant* keywords. One of the widely used methods in document representation is called the *TFIDF* (Term Frequency Inverse Document Frequency) method that weighs the significance of a keyword, based on its overall frequency in a document set (Term Frequency) and the number of documents that have at least one instance of a given keyword. The two components of the TFIDF metric are computed over the entire corpus and significant keywords are then selected. High value TFIDF terms basically indicate terms that appear to be significant for the specific document. The more frequent a term is in a document and the less it appears in other documents the higher its weight. A term that appears in every other document has a zero weight. TFIDF weight ($tfidf_{i,j}$) of a token i in document j is given as [230]:

$$tfidf_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \tag{5.2}$$

where $tf_{i,j}$ is token frequency of token i in document j, N represents number of documents in a collection, and df_i is the document frequency of token i in the collection.

Figure 5.1 represents the steps followed for TFIDF text feature extraction. The first step is to extract the text from the textual files and convert them into text strings, and then the numbers (if there are any) are removed from the strings. These strings are used to create tokenized documents which are then pre-processed by lemmatization and removing punctuation marks, stop words, and words with length 1 or 2. Afterward, a cleaned bag of words is created from the cleaned documents, and this bag is used for TFIDF feature extraction.

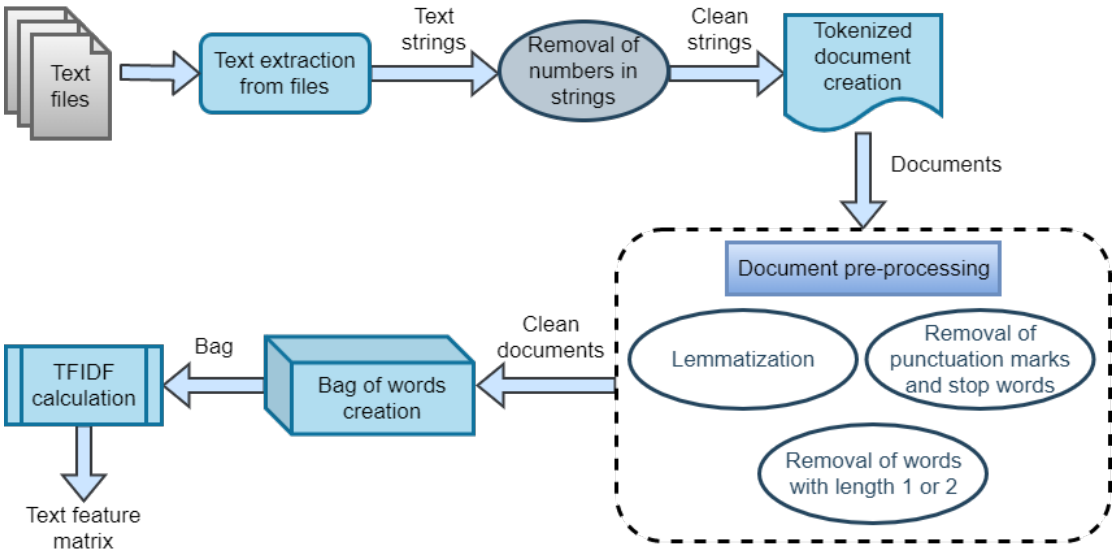


Figure 5.1: Process flow for TFIDF text feature extraction

5.4 Hebbian and Oja learning rule

The neuropsychologist *Donald Hebb* postulated regarding the learning of the biological neurons [193]:

"When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place on one or both cells such that A's efficiency as one of the cells firing B, is increased."

It can be stated that if two neurons, which are connected, are activated simultaneously on some input, then the connection is strengthened. In simple terms, "Neurons that fire together, wire together". In the basic formulation, the simple Hebbian learning depends only on the presynaptic a_i and postsynaptic a_j firing rate and a learning rate η.

$$\Delta w_{ij} = \eta \cdot a_i a_j \tag{5.3}$$

where Δw_{ij} represents the change in weight of synapse connecting i^{th} neuron with j^{th} neuron. Hebbian learning is a correlation-based learning principle.

Let the postsynaptic activity value over multiple input synapses is evaluated by:

$$a_j = \sum_i w_{ij} a_i = \mathbf{a}^T \times \mathbf{w}_j \tag{5.4}$$

the learning rule cumulates the auto-correlation matrix $\mathbf{Q}$ of the input $\mathbf{r}$:

$$\Delta \mathbf{w}_j = \eta \mathbf{a} a_j = \eta \mathbf{a} \times \mathbf{a}^T \times \mathbf{w}_j = \eta Q \times \mathbf{w}_j \tag{5.5}$$

Q depicts the correlation matrix of the inputs when several input vectors are introduced:

$$Q = \mathbb{E}_{\mathbf{a}}[\mathbf{a} \times \mathbf{a}^T] \tag{5.6}$$

Hence, Hebbian plasticity is learning strong weights to frequently co-occurring input elements. The simple Hebbian learning rule suffers from a severe issue. There is nothing to stop the connections from growing all the time, eventually leading to huge values. So, weights will keep on growing in size with time. Another term is required to balance this growth. A term depicting "forgetting" has been utilized in several neuron models where the weight value itself should be subtracted from the right-hand side. *Erkki Oja*, a *Finnish computer scientist* proposed a learning rule, known as *Oja rule*, which is a mathematical formalization of the Hebbian rule, such that a neuron learns to compute a principal component of its input stream over time [231]. The main idea behind this rule is to make the forgetting term proportional to the value of weight along with the square of the activity of the postsynaptic neuron. It can also be called as a normalized form of the Hebbian learning rule. Oja rule normalizes the length of a weight vector by a local operation:

$$\Delta w_{ij} = \eta a_i a_j - \eta a_j^2 w_{ij} \tag{5.7}$$

$\eta a_j^2 w_{ij}$ is a regularization term. When the postsynaptic activity a_j or weight w_{ij} are too large, then the term cancels the Hebbian $a_i a_j$ part and decreases the weight. The new (next) weight can be calculated as the old weight ($w_{ij}(n-1)$) plus the change in weight (Δw_{ij}) as given in eq. (5.8):

$$w_{ij}(n) = w_{ij}(n-1) + \Delta w_{ij} \tag{5.8}$$

5.5 Oja learning based cross-modal retrieval using deep features

5.5.1 Problem formulation

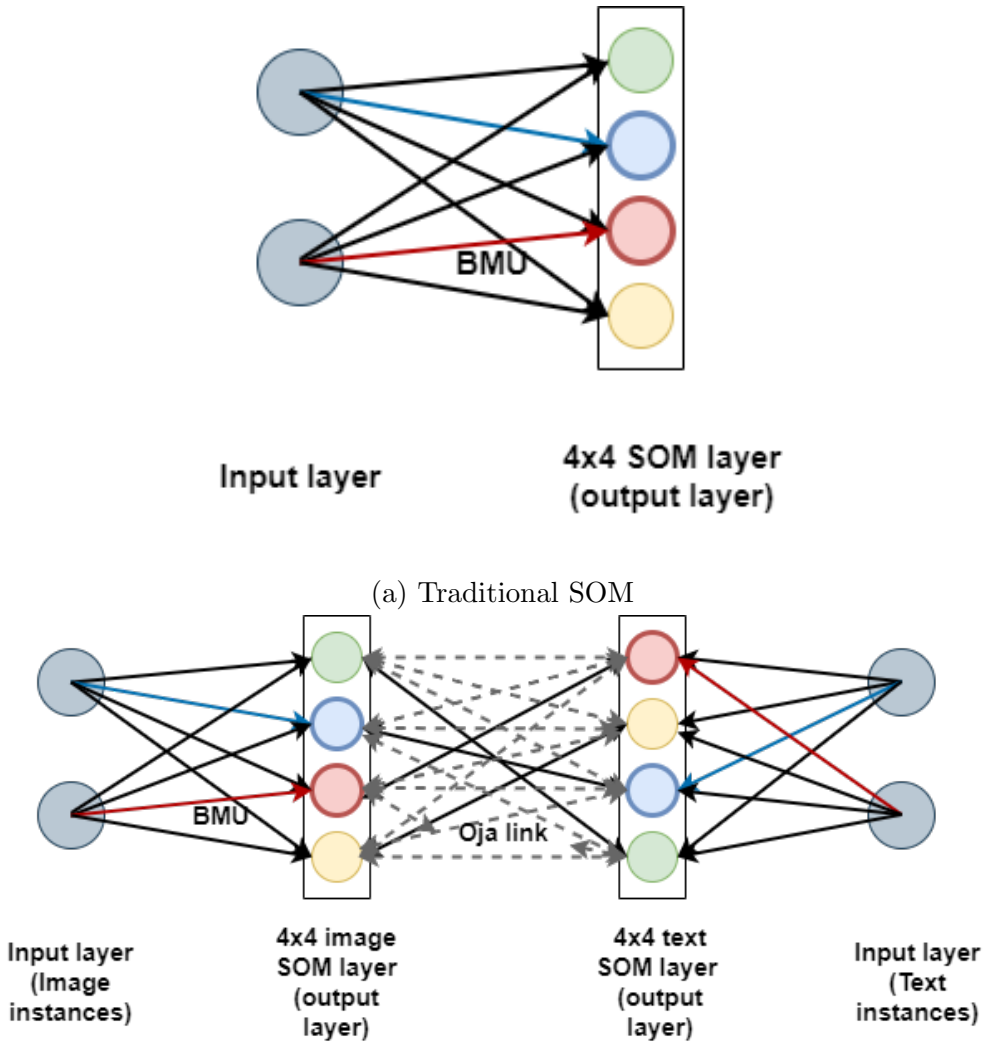
The goal is to make a robust connection between strongly related gastrointestinal images and collateral text by reducing the semantic gap between these hetero-

geneous modalities as much as possible. We have a collection of images and corresponding text in the form of multiple labels (representing the images). Each image file has a single text file related to it. The aim of the proposed approach is to retrieve the matched images or text given a text or an image query respectively. Let $E = (I_j, T_j, L_j)_{j=1}^{N}$ depicts the total endoscopy image-text dataset, where $I_j \in R^{d_I}$ and $T_j \in R^{d_T}$ represents the image and text features having different dimensions d_I and d_T, respectively. (I_j, T_j) is an image-text pair with same semantic label $L_j \in R^c$, where c symbolizes the total number of categories of semantic concepts in the dataset. The labels are not used in the whole model training process as the proposed technique is of unsupervised nature, however, they have been used while evaluation of the performance metric for the trained cross-modal system. Total image-text pair instances N have been divided into N_1 training instances and N_2 testing instances, creating the train data $E_{train} = (I_k, T_k)_{k=1}^{N_1}$ and test data $E_{test} = (I_k, T_k)_{k=1}^{N_2}$. The text training set is defined as $T_{train} = [T_1, T_2, ..., T_{N_1-1}, T_{N_1}] \in R^{d_T \times N_1}$ and image training set as $I_{train} = [I_1, I_2, ..., I_{N_1-1}, I_{N_1}] \in R^{d_I \times N_1}$. Similarly, $I_{test} = [I_1, I_2, ..., I_{N_2-1}, I_{N_2}] \in R^{d_I \times N_2}$ depicts the image testing set along with $T_{test} = [T_1, T_2, ..., T_{N_2-1}, T_{N_2}] \in R^{d_T \times N_2}$ as the text testing set.

5.5.2 Proposed technique

A detailed description of the traditional self-organizing map has been given in the chapter (3). The proposed approach aims to associate two separately trained traditional self-organizing maps (on diverse modalities) using improved Hebb links or Oja links, creating a hybrid SOM model which can be utilized for cross-modal image-text retrieval. Figure 5.2 demonstrates the difference between the traditional SOM (Figure a) and hybrid SOM (HSOM) (Figure b) using two data instances in the input layer. HSOM associates two trained traditional SOMs using Oja links, as shown in the figure.

So in the proposed study, one SOM is dedicated to the image modality (dubbed image SOM) and the other to the text modality (dubbed text SOM). These two SOMs are integrated using a third network known as the improved Hebbian network or Oja network that connects each neuron (similar cluster of images) in image SOM with every neuron (similar cluster of texts) in the text SOM. Oja network is inspired by the Oja learning rule described in *Section (5.4)*. Neurons in the trained SOMs that are synchronously highly active while training are associated via the Oja network. This network has been used for the association to enhance the connections between the SOMs when respective nodes in them activate in response to an input text or an image. If the size of the text SOM is $m \times n$ and



(a) Traditional SOM

(b) Hybrid SOM

Figure 5.2: Demonstration of (a) traditional SOM; and (b) hybrid SOM where two separately trained traditional SOMs are integrated using Oja links.

image SOM is $p \times q$, then the size of the connecting network would be $m \times n \times p \times q$. In the proposed study, the size of both the image and text SOM are 4×4, so the size of the Oja network is 16×16.

For implementing the proposed approach, image and text features are retrieved as described in the sections (5.2.1, 5.3). Two independent SOMs net_T and net_I of dimension 4×4 are trained for texts and images correspondingly, and the SOM neuron numbers are also obtained corresponding to each instance depicted as $classes_T$ and $classes_I$ matrices. The SOM node weights represented by $nodeWeights_I$ and $nodeWeights_T$ are obtained for both image and text SOM for further experimentation. Let $winnersMatrix_T$ and $winnersMatrix_I$ be the weight vector of the winner node of each text and image input instance, respectively. Afterward, Euclidean distance has been calculated between each input vector and the corresponding winner node weight vector, and the results are depicted as one-

dimensional matrices $woja_I$ and $woja_T$ individually. Now the training of the Oja network is performed as per Eq. 5.9, and the Oja link weights (represented by $ojaLink$ matrix) keep on updating for each input instance for the whole training process. All the nodes of net_I are connected with all the nodes of net_T in the Oja network. However, the strength of the Oja bond is determined by the Oja link weight.

$$\begin{aligned} ojaLink(classes_I(i), classes_T(i)) = ojaLink(classes_I(i), classes_T(i)) \\ +(\alpha * woja_I(i) * woja_T(i) \\ -\alpha * woja_T(i) * ojaLink(classes_I(i), classes_T(i)) * woja_T(i)); \end{aligned} \tag{5.9}$$

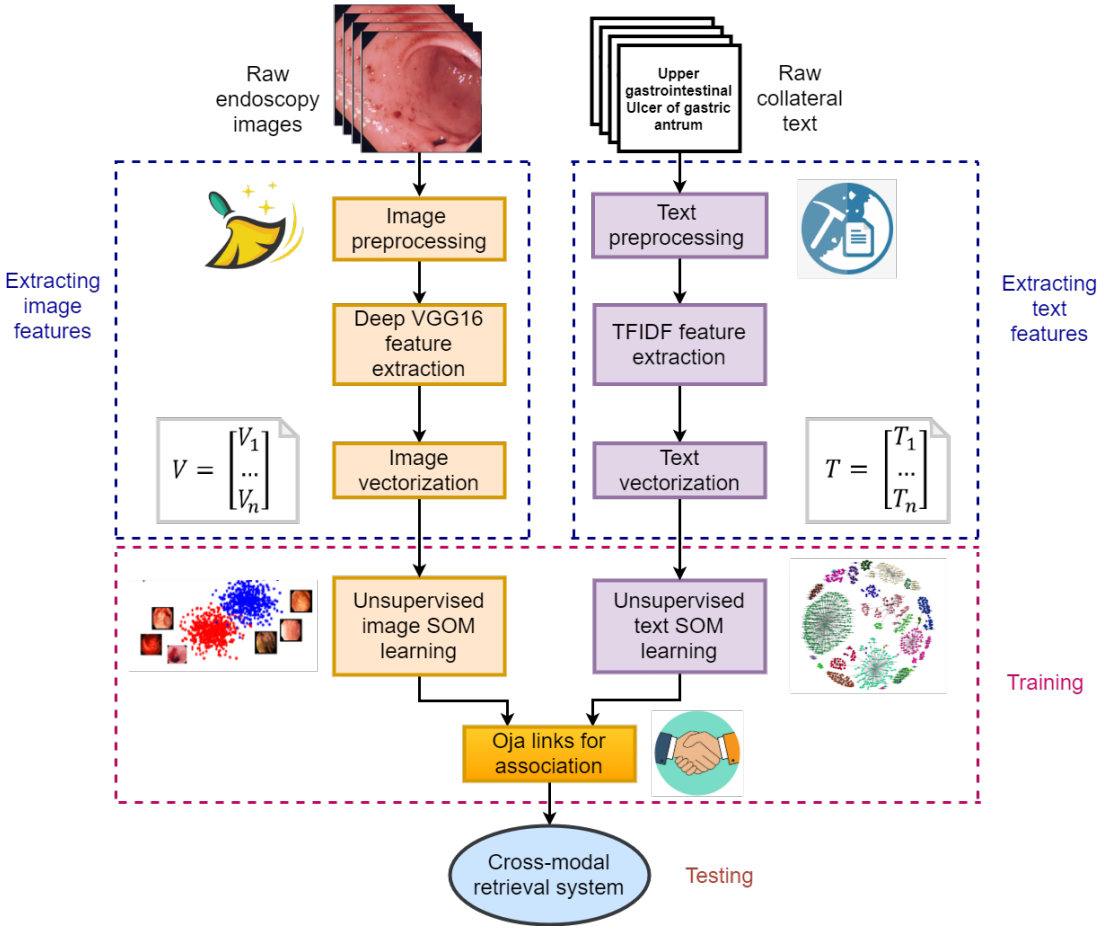


Figure 5.3: Process flow of image and text training for the proposed hybrid cross-modal retrieval system

where $1 \leq i \leq length(classes_I)$ and the α represents the learning rate whose value has been chosen to be 0.001 after experimentation on values given in the set $\{0.1, 0.01, 0.001\}$. After the network training, two vectors $Anet_I$ and $Anet_T$ of size 16 are created such that $Anet_I$ will have the node numbers of net_T having the highest Oja link weight where each index of the $Anet_I$ vector represent the

Algorithm 5 Algorithm of HSOM technique exploiting Oja rule for endoscopy cross-modal retrieval

INPUT: E_{train} and E_{test}

OUTPUT: Trained net_I and net_T SOMs, retrieval of matched images and text corresponding to text and images in E_{test}, respectively

procedure IMAGE FEATURE EXTRACTION
  Input all images
  Resize the images to $224 \times 224 \times 3$ ▷ Defined image input size for VGG16 input
  Extract the deep features of 1000 dimension from *fc8* layer of VGG16 network
end procedure
procedure TEXT FEATURE EXTRACTION
  Input all text files
  Removal of numbers from each text
  $cleanedDocuments \leftarrow tokenizedDocument(text)$ ▷ Create tokenized documents from the text
  Perform lemmatization
  Remove punctuation marks, stop words and words with $length \leq 2$
  $cleanedBag \leftarrow bagOfWords(cleanedDocuments)$ ▷ Create a bag-of-words from cleaned documents
  Calculate TFIDF score from the cleaned bag using eq. (5.2)
end procedure
procedure HSOM BASED CROSS-MODAL RETRIEVAL USING ENDOSCOPY DATA
  Load E_{train} and E_{test}
  $dimension1 \leftarrow 4, dimension2 \leftarrow 4$ ▷ Dimensions of both image and text SOM
  $net_I \leftarrow selforgmap([dimension1 dimension2])$ ▷ Configure image SOM with default parameters except dimensions
  $net_T \leftarrow selforgmap([dimension1 dimension2])$ ▷ Configure text SOM with default parameters except dimensions
  $net_I \leftarrow train(net_I, I_{train}), net_T \leftarrow train(net_T, T_{train})$ ▷ Training of maps
  $classes_I \leftarrow vec2ind(net_I(I_{train})), classes_T \leftarrow vec2ind(net_T(T_{train}))$ ▷ Retrieving node number for each input instance
  for $i \leftarrow 1$ `to` $length(classes_I)$ **do** ▷ Winner node weight matrix corresponding to image input instances
    $winner_I \leftarrow classes_I(i)$
    $winnersMatrix_I(:, i) \leftarrow nodeWeights_I(winner_I, :)'$
  end for
  for $i \leftarrow 1$ `to` $length(classes_T)$ **do** ▷ Winner node weight matrix corresponding to text input instances
    $winner_T \leftarrow classes_T(i)$
    $winnersMatrix_T(:, i) \leftarrow nodeWeights_T(winner_T, :)'$
  end for
  for $i \leftarrow 1$ to $length(classes_I)$ **do** ▷ Euclindean distance calculation
    for $j \leftarrow 1$ to $imageVectorDimension$ **do**
      $woja_I(i) \leftarrow woja_I(i) + (winnersMatrix_I(j, i) - input_I(j, i))^2$
    end for
    for $j \leftarrow 1$ to $textVectorDimension$ **do**
      $woja_T(i) \leftarrow woja_T(i) + (winnersMatrix_T(j, i) - input_T(j, i))^2$
    end for
    $woja_I(i) \leftarrow sqrt(woja_I(i))$
    $woja_T(i) \leftarrow sqrt(woja_T(i))$
  end for
  for $i \leftarrow 1$ to $length(classes_I)$ **do**
    Train the improved Hebbian (Oja) network using Equation 5.9
  end for
  Follow Algorithm 2 for creation of $Anet_I$ and $Anet_T$
  Cluster $I_k \in net_I$ and $T_k \in net_T$ where $(I_k, T_k) \in E_{test}$ and $k \in [1, N_2]$
  Refer $Anet_I$ and $Anet_T$ to find the corresponding Oja link node
  Retrieve results from the *found* node
end procedure

node number in net_I. Similarly, $Anet_T$ comprises the node numbers of net_I SOM. For model testing using an image input, firstly, the input image is clustered in a suitable node in the image SOM as per the similarity. Afterward, the corresponding linked node in the text SOM is found, and the clustered instances from both the image and text nodes are retrieved. The Algorithm 5 displays all the steps performed for implementing the proposed method. In this algorithm, the steps required for creating $Anet_I$ and $Anet_T$ are similar to the steps mentioned in the Algorithm 2. Figure 5.3 shows the pipeline diagram of the proposed system for

cross-modal gastrointestinal images and text retrieval.

5.6 Conclusion

This chapter introduced new ways of indexing and querying image collections by training neural computing systems with images accompanied by collateral texts. The characteristic visual features of the image collection are extracted using pre-trained VGG16 deep convolution neural network. The characteristic linguistic features of the collateral texts are extracted using TFIDF, a well-known text representation method. The images and keywords were categorized synchronously but separately, by using an unsupervised clustering algorithm – Kohonen self-organizing feature maps. SOMs learn to categorize unseen images and collateral texts and concurrently, during uni-modal learning, an Oja link is established between the most active nodes in the two uni-modal maps: This is the basis of our claim that we use multi-modal features to train neural networks and during the training to establish cross-modal connections between the two maps through another unsupervised network, the improved Hebbian learning network or Oja network.

Chapter 6
Experimental Analysis

This chapter provides all the experimentation details, such as parameter settings, evaluation metrics, dataset description, and model training, along with the obtained results after the experiments. The experiments have been performed on two datasets: the public Wikipedia dataset and the primary endoscopy dataset. The results obtained (using the MAP score evaluation metric) on the Wikipedia dataset have been compared with traditional and state-of-the-art cross-modal techniques, and the proposed hybrid SOM approach outperforms them. Two operations have been performed independently on the primary endoscopy data: image clustering and image-text cross-modal retrieval. Image clustering is performed using a traditional self-organizing map. Cross-modal retrieval has been performed using VGG16 features for visual representation and TFIDF features for text representation. Two SOMs are trained separately using these features and then associated using Oja links. The results of this proposed approach have also been compared with the technique proposed for Wikipedia cross-modal retrieval. All the experiments have been performed in *MATLAB R2019a*.

6.1 Case study on public Wikipedia dataset

6.1.1 Dataset

The proposed approach has been tested on *Wikipedia*[1] [43] dataset which includes a document corpus consisting of linked image and text pairs. It has been composed of Wikipedia's "featured articles" which accompanied by one or more images from Wikipedia Commons, giving a pair of appropriate variety. Articles are categorized into 29 categories by Wikipedia with an individual categorization of image and text elements. Only the top 10 bulky categories are considered by most of the researchers for experimentation as the remaining categories have scarce data. The final data corpus classified into 10 semantic categories contain 2,866 documents in total. It has been arbitrarily bifurcated into a training and testing set comprising of 2,173 and 693 documents respectively. Division of each class' documents in the train and test set is presented in Table 6.1.

[1] `http://www.svcl.ucsd.edu/projects/crossmodal/`

Table 6.1: Train and test split of Wikipedia classes

Class	Train	Test	Total
History	248	85	333
Art & architecture	138	34	172
Media	178	58	236
Biology	272	88	360
Royalty & nobility	144	41	185
Geography & places	244	96	340
Warfare	347	104	451
Literature & theatre	202	65	267
Music	186	51	237
Sport & recreation	214	71	285

6.1.2 Evaluation metrics

The commonly used metric for detecting the efficiency of a cross-modal retrieval method is Mean Average Precision (MAP). It tests whether the obtained outcome belongs to the same category as query (relevant) or not (irrelevant) [73]. It is the mean of the measured average precision (AP) across all the queries. Provided a query (a text or an image) and a set of respective retrieved outcomes Y, AP can be evaluated as:

$$AP = \frac{1}{R} \sum_{y=1}^{Y} P(y) rel(y) \tag{6.1}$$

where R represents the ground truth positives or the number of relevant results in the retrieved results [232], $P(y)$ depicts the precision of top y retrieved results, if the y^{th} retrieved result is relevant then $rel(y) = 1$ and otherwise 0. Now, MAP can be calculated as:

$$MAP = \frac{1}{N} \sum_{n=1}^{N} AP \tag{6.2}$$

where N represents number of queries. The more is the MAP score value, the better the algorithm is.

6.1.3 Comparison methods

Following points provide the brief description about the methods used for comparison with the proposed *HSOM* approach. Table 6.2 presents the characteristics of these comparison techniques.

1. *CCA* [43] is a fundamental subspace learning based method. It finds the pair of projections for different modes so that the relation between them is augmented.

2. *SM* [43] represents the multi-modal data at an upper level of abstraction such that there is a natural correspondence between diverse modality spaces. Moreover, it utilizes multi-concept logistic regression for the classification of both text and image modalities.

3. *SCM* [43] is the amalgamation of CCA and SM. Firstly, it uses CCA for attaining feature representations and then utilize these representations in building a semantic space.

4. *DDL* [233] is a scalable hierarchical learning framework that deals with weakly paired diversified data. In the learned representation space for using label knowledge, a shared classifier is applied across diverse modalities. A modal invariant representation is achieved by enforcing low-rank constraint across modalities.

5. *DLA-CMR* [66] is an adversarial cross-modal retrieval method that is based upon dictionary learning. Adversarial learning extracts each modality's numerical attributes, whereas dictionary learning functions as a feature reconstructor to reconstruct distinguishing features.

   Adversarial learning extracts the statistical attributes of each modality whereas dictionary learning act as a feature re-constructor for reconstructing discriminative features.

6. *DAML* [91] maps classified multi-modal data pairs onto a shared latent feature subspace in a nonlinear fashion. This augments the inter-class variation and reduces the intra-class variation and the divergence of each data pair obtained from two modes of the same concept. An additional regularization is added by the introduction of adversarial learning.

7. *SCCMR* [54] combines the label prediction and projection matrices' optimization into an integrated framework for achieving a globally optimum result. Graph embedding is utilized in this for considering nearby neighbors

Table 6.2: Characteristics of the compared methods. U = Unsupervised, S = Supervised and Se = Semi-supervised

Characteristics/ Methods	**Type**	**Subspace learning**	**Graph regularization**	**Semantic information**	**Inter-class and intra-class correlation/ similarity**	**Dictionary learning**	**Adversarial learning**	**Deep learning based**
CCA [43]	U	✓						
SM [43]	S			✓				
SCM [43]	S	✓		✓				
DDL [233]	-	✓		✓		✓		✓
DLA-CMR [66]	-			✓	✓	✓	✓	
DAML [91]	S	✓			✓		✓	✓
SCCMR [54]	Se	✓	✓	✓	✓			
CSGL [55]	S	✓	✓	✓				
SGRCR [56]	S	✓	✓		✓			
CCDCR [57]	S		✓	✓	✓			
CR-CDSL [94]	S	✓		✓				✓
TQSL [104]	S	✓	✓	✓	✓			
KDM [105]	S	✓		✓	✓			
MJSL [106]	S	✓		✓				
SMDCR [107]	Se	✓		✓				

in the potential subspace of paired text and images and text and images with identical semantics.

8. *CSGL* [55]makes use of semantic information and learns the projection matrix in integration rather than separately for each modality for more discriminative projection. It considers the consistency among diverse modalities by incorporating graph regularization for conserving the organization of original data in the projective space. A collaborative learning scheme is utilized for avoiding suboptimal solution and integration of diverse modalities for better projection.

9. *SGRCR* [56] learns two couple of projections as per diverse retrieval tasks. It projects the diverse modal data onto an isomorphic common subspace and heterogeneous adjacent graphs are built for conserving the correlation among different modalities. It considers the inter and intra class similarity of modalities in an integrated framework. Feature selection is performed by L_2 norm.

10. *CCDCR* [57] minimizes the inter-modality distance and maximizes the intra-modality distance of class center samples for reinforcing the discriminative capability of the model. In order to further enhance semantic similarity between different modalities, a multi-modal graph consisting of an inter-modality similarity graph, class center inter-modality graph, and intra-modality graph is fused into the technique. This approach considers both local as well as global structural information of data.

11. *CR-CDSL* [94] exploits the latent semantics of untagged multi-modal information with joint deep semantic learning for increasing the discerning ability of supervised retrieval model. For mutually projecting image and text samples into a common semantic representation, two corresponding neural networks are trained. Weak semantic labels of both unlabeled text and images are produced consequently based on them. They are mutually exploited with categorized training samples for retraining the model which eventually finds a more semantically meaningful subspace for better cross-modal retrieval.

12. *TQSL* [104] is a subspace learning approach which is dependent on task as well as query. It is an integrated cross-modal framework where class and task-specified subspaces are learned together using an effective iterative optimization and a task-category-projection mapping table is created based on it. A semantic mapping function is learned between multi-modal documents and corresponding classes by a trained linear classifier.

13. *KDM* [105] is a cross-modal subspace learning framework which is based upon correlation. Unlike most other methods that directly maximize feature correlations across multi-modal data, it learns subspace representation for each modality by augmenting the kernel dependency. This approach maps the modalities into diverse Hilbert spaces with the same dimension separately. Afterward, the kernel matrix is calculated in each Hilbert space and the correlations are measured across modalities on the basis of kernels.

14. *MJSL* [106] mines the latent common knowledge of semantic overlap to the maximum degree possible. It selects the high-level semantics, keeps the pairwise closeness, and selects the appropriate features for attaining the most discriminative subspace for each modality.

15. *SMDCR* [107] is a semi-supervised cross-modal technique which is modality-dependent and uses both labeled and unlabeled samples for getting two couple of projection matrices. It utilizes the feature distance for representing the semantic knowledge of unlabeled samples in the optimization process for getting full use of data structural information. It fully utilizes the semantic knowledge of whole multi-modal data and data distribution property.

6.1.4 Parameter settings

The values of certain parameters in the implementation have been chosen such that the overall performance is increased. ZM features are extracted at order 5, so the total retrieved features are 12 for each image instance. These features have the least redundancy due to the orthogonal characteristics of moments. For setting the appropriate value of the total number of topics in the LDA model for text feature extraction, perplexity and time analysis has been performed. *Perplexity* is the statistical measure of how well a sample is predicted by a probability model. The aim is to choose the number of topics that minimize the perplexity value. Moreover, with an increase in the number of topics, the LDA model may take more time to converge. So to handle this trade-off, both the values have been plotted simultaneously for the different number of topics as shown in Figure 6.1. It can be concluded from the figure that 14 is a good choice for the total number of topics. So, it has been chosen in the final LDA model.

6.1.5 Model training

Figure 6.2 illustrates the train data distribution after individual image and text SOM training. The data is evenly distributed among the SOM nodes. These results are obtained with 12-d ZM visual features and 14-d LDA linguistic features.

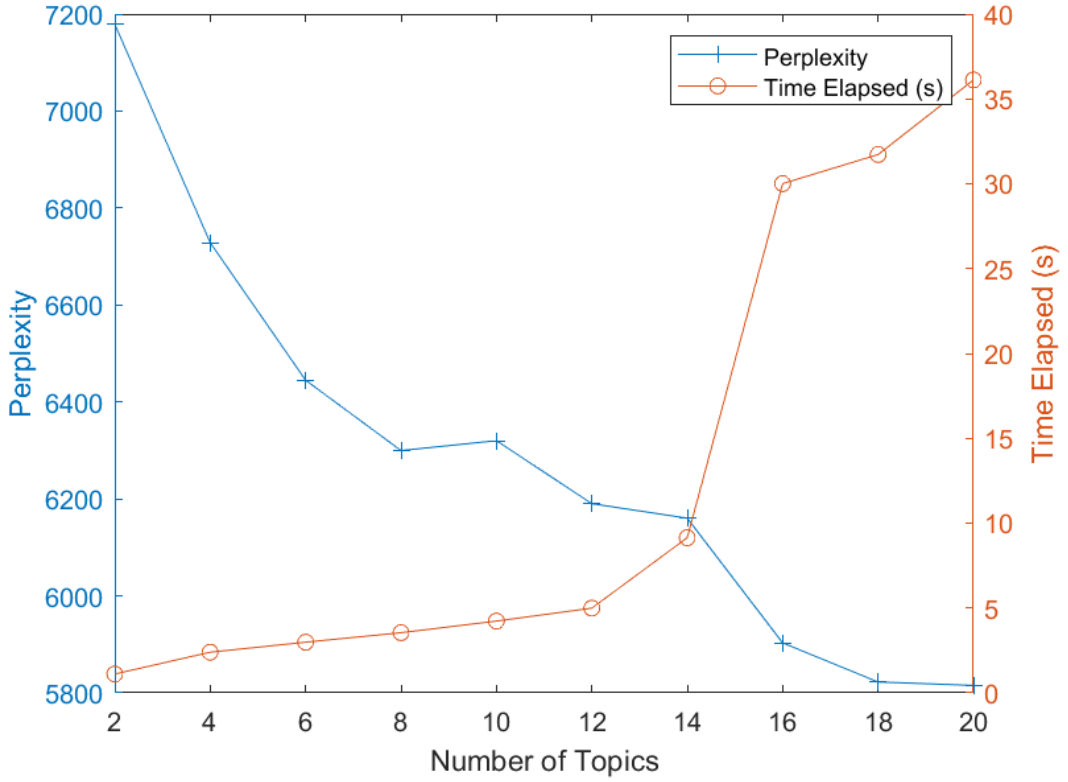
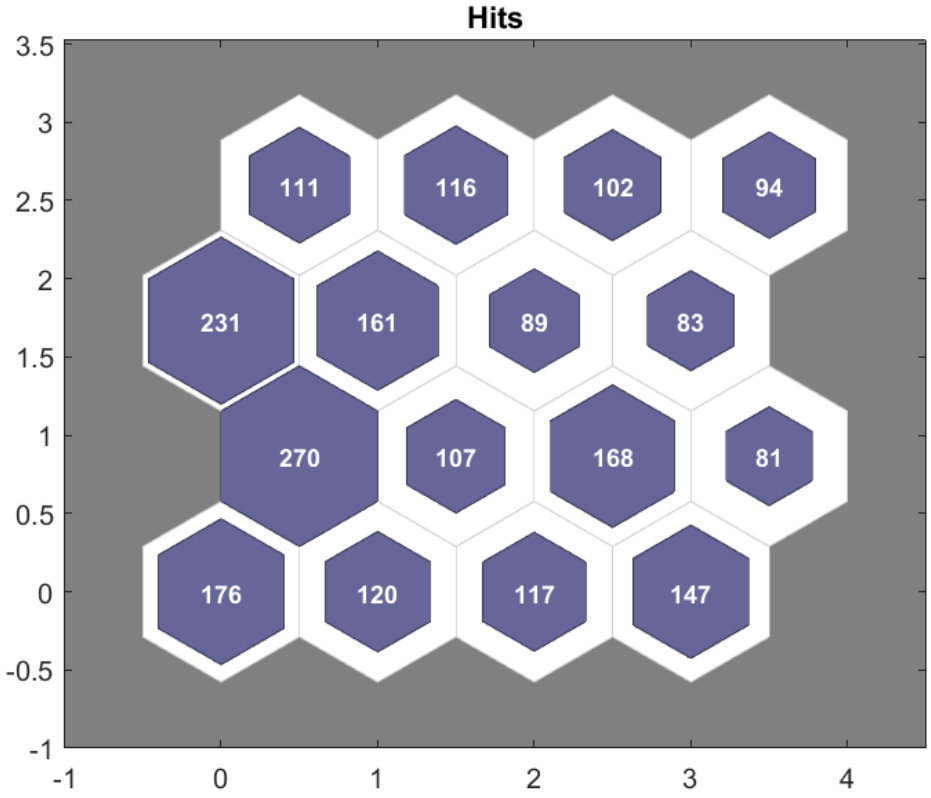
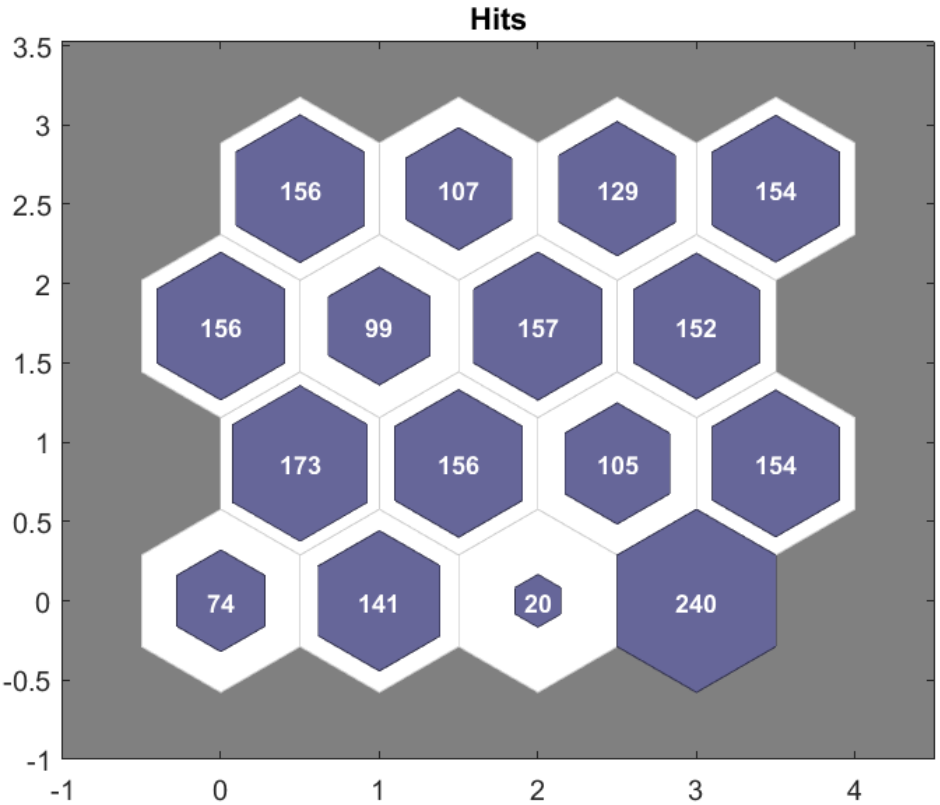


Figure 6.1: Perplexity and time analysis for choosing an appropriate number of topics for the LDA model

Figure 6.3 demonstrates the distances between the neighboring nodes for the respective image and text SOM. Nodes are represented by blue hexagons which are connected to their neighbors using red lines. The colors in the red line sections depict the distance between nodes. The darker the color, the more is the distance. A band of dark sections traverses from the bottom centre region to the middle right region making a reversed 'L' shape in the image SOM (Figure 6.3a). It appears that the clusters of images have been divided into two sets such as the clusters in the lower right corner and the rest of the SOM clusters. However, the third node at the bottom of the text SOM (Figure 6.3b) is lying at a huge distance from all the other nodes and it seems to be acting as an outlier. In the corresponding trained SOM Figure 6.2b, the same node has the least number of instances in comparison to all other SOM nodes. The positions of weight vectors and data points are displayed in Figure 6.4. The appropriate learning rate for Hebbian network training is chosen after analyzing the average image-text query MAP score by training multiple times at diverse values such as $[0.001, 0.005, 0.01, 0.05, 0.1, 1]$. From Figure 6.5, it is noticeable that the 0.1 learning rate value is suitable for the experimental analysis.

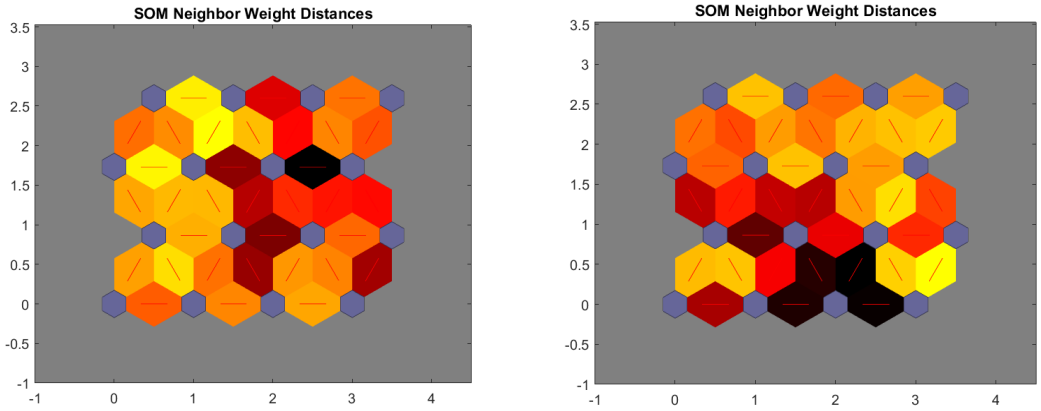


(a) Trained image SOM



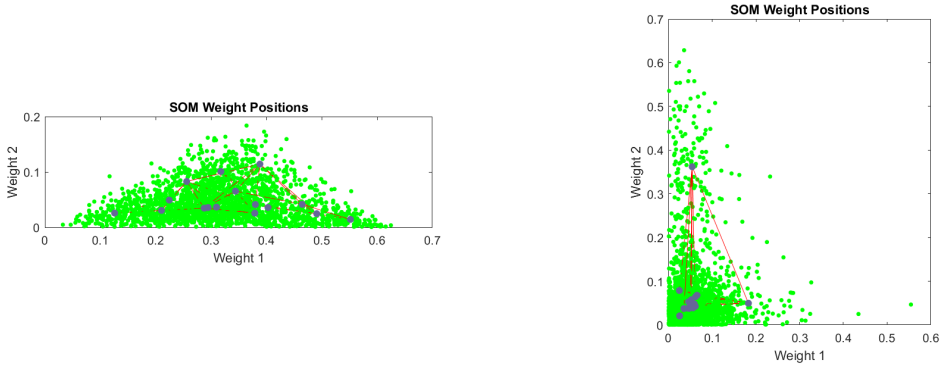
(b) Trained text SOM

Figure 6.2: Input train data distribution after individual SOM training



(a) Trained image SOM (b) Trained text SOM

Figure 6.3: Neighbor Distances among respective SOM nodes after training. Darker shade denotes larger distance.



(a) Trained image SOM (b) Trained text SOM

Figure 6.4: Location of data points and weight vectors

6.1.6 Results

Table 6.3 demonstrates the comparison of various state-of-the-art methods with the proposed technique on the basis of MAP score where *I2T* means retrieving collateral text using an image query and *T2I* means retrieving matched images using a textual query. *Average* denotes the average MAP score for I2T and T2I experiments. The column *Feature type* depicts the type of image and text features utilized in that particular method where **H** designates handcrafted features (128-d visual SIFT representation for images, 10-d LDA representation for text) and **D** stands for Deep features (4096-d CNN for images, 100-d LDA for text).

The handcrafted features are utilized by authors in [44] for representing images and collateral text. These features are freely provided by the authors on the *link*[1] along with the Wikipedia dataset. The base representation for both images and text is Bag-of-Words (BOW). Firstly, a bag of SIFT features is extracted per

[1] `http://www.svcl.ucsd.edu/projects/crossmodal/`

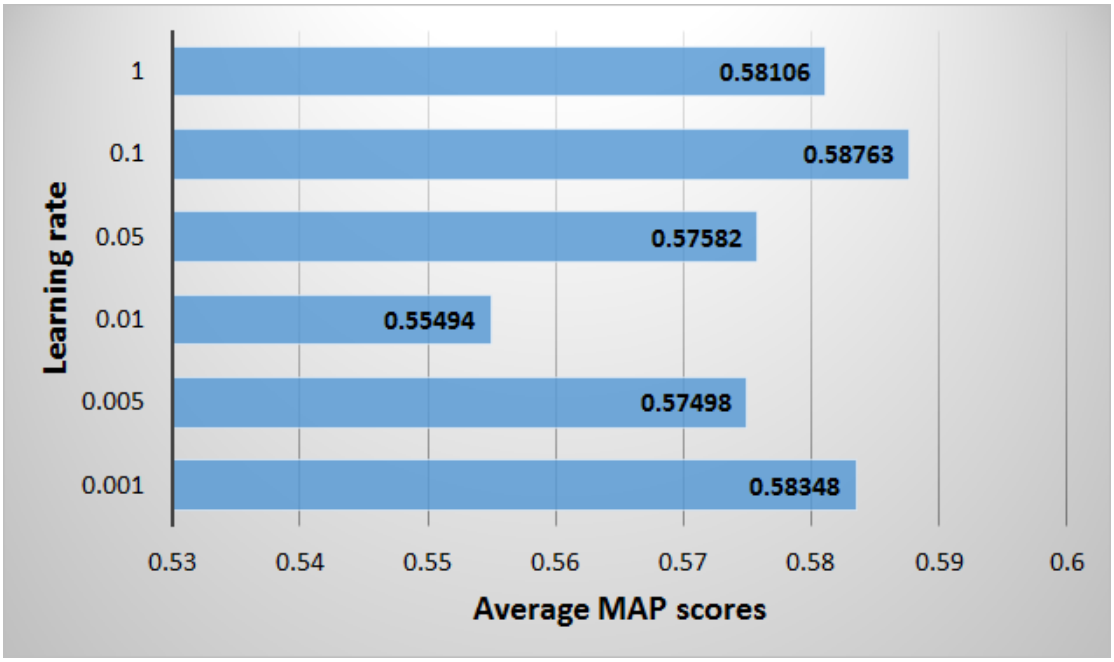


Figure 6.5: Average MAP scores at different values of learning rate for training Hebbian network

training image[1] and a visual word codebook is learned using K-means clustering. Afterward, SIFT descriptors are vector quantized with the codebook to create a visual word counts vector. Text words that are obtained by stemming the text with the Python Natural Language Toolkit[2], are fit by LDA model [172] by utilizing the implementation of [234]. Almost all the researchers who have tested their respective cross-modal methods on the Wikipedia dataset have compared the MAP score results by utilizing these features as well. Hence, they have been considered in this study also for comparative analysis.

In the table, SL_HSOM represents the results obtained using handcrafted features, and ZL_HSOM depicts the results achieved using 12-d ZM for images and 14-d LDA for text. It is evident from the table that the ZL_HSOM approach is better than the other methods and so the MAP scores are highlighted. Figure 6.6 shows the performance chart of all the methods based on their MAP scores and Figure 6.7 presents the class-wise performance of the proposed technique. Table 6.4 shows the I2T, T2I and their average MAP score values for respective dataset categories. Figure 6.9 demonstrates a curve depicting the precision values obtained for each test query (image in case of I2T and text in case of T2I operation) in a sorted manner and the change in precision values as per the queries can be visualized. Figure 6.8 illustrates a few matched images and text results retrieved using an image query on trained ZL_HSOM model.

[1] https://lear.inrialpes.fr/people/dorko/downloads.html
[2] http://www.nltk.org/

Table 6.3: MAP comparison of prominent recent methods with proposed approach on Wikipedia dataset. **H** means handcrafted features and **D** represents deep features.

Method	**MAP Score**			**Feature type**
	I2T	T2I	Average	
CCA [43]	0.249	0.196	0.223	H
SM [43]	0.225	0.223	0.224	H
SCM [43]	0.277	0.226	0.252	H
DDL [233]	0.2832	0.2615	0.2724	H
	0.3812	0.3501	0.3657	D
DLA-CMR [66]	0.369	0.261	0.315	H
	0.539	0.453	0.496	D
DAML [91]	0.356	0.267	0.322	H
	0.559	0.481	0.52	D
SCCMR [54]	0.431	0.403	0.417	D
CSGL [55]	0.3996	0.3904	0.395	H
SGRCR [56]	0.284	0.227	0.2555	H
	0.4365	0.406	0.421	D
CCDCR [57]	0.2849	0.2253	0.2551	H
CR-CDSL [94]	0.348	0.249	0.299	H
	0.508	0.442	0.475	D
TQSL [104]	0.463	0.415	0.439	D
KDM [105]	0.4562	0.4785	0.4674	D
MJSL [106]	0.4432	0.3832	0.4132	D
SMDCR [107]	0.284	0.232	0.258	H
	0.43	0.428	0.429	D
SL_HSOM	0.5872	0.4744	0.5308	H
ZL_HSOM	**0.6461**	**0.5228**	**0.5844**	-

SOM vs HSOM

This section demonstrates the comparison of results obtained using traditional SOM and HSOM. In the traditional SOM implementation, all the required parameter values, visual and textual feature vectors are the same as considered for

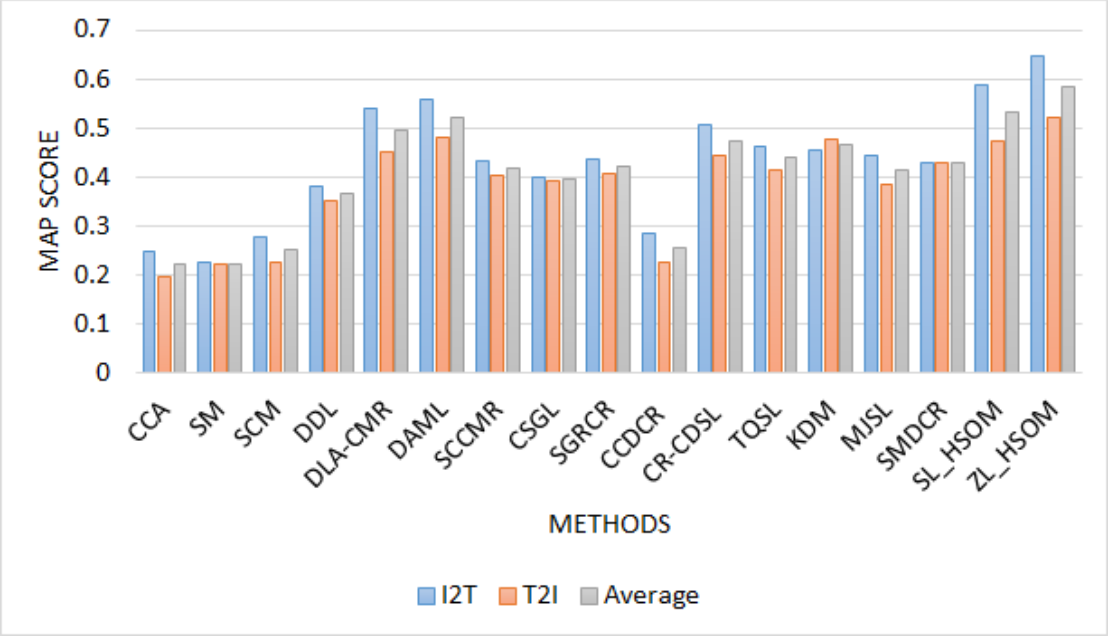


Figure 6.6: Performance analysis based on MAP scores

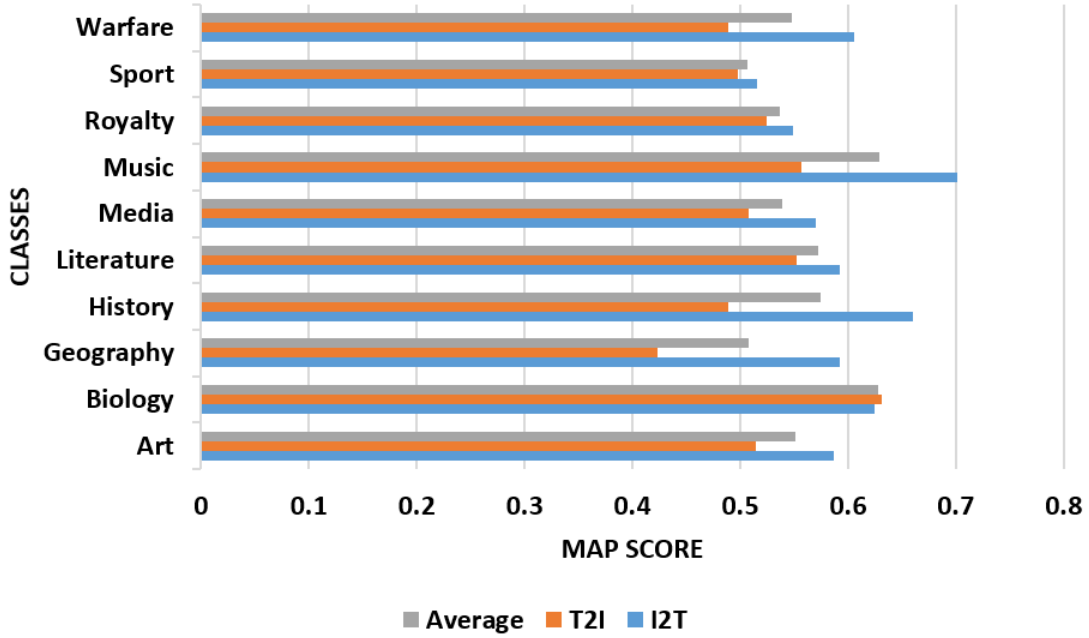


Figure 6.7: Performance chart based on MAP scores for each class

HSOM implementation. The MAP score is evaluated for retrieval of related images using an image query (*I2I*) and retrieval of related text using a textual query (*T2T*). Table 6.5 shows the comparison of MAP score values obtained in the tasks I2I, T2T, I2T, and T2I using diverse visual and textual features as explained in the above sections. It is evident from the table that the cross-modal retrieval has high MAP score than uni-modal retrieval and hence information from multiple sources (such as image and text together) always results in better performance

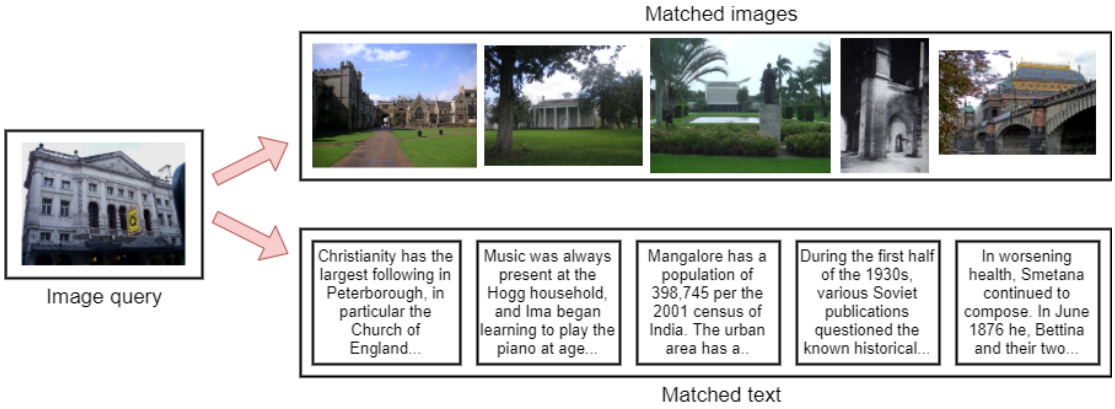


Figure 6.8: Retrieved image and text results using an image query

Table 6.4: Category-wise MAP scores based on proposed technique

Categories	I2T	T2I	Average
Art	0.5867	0.515	0.5509
Biology	0.6244	0.6314	0.6279
Geography	0.5923	0.423	0.5077
History	0.6606	0.4885	0.5746
Literature	0.592	0.5524	0.5722
Media	0.5706	0.5078	0.5392
Music	0.7012	0.5571	0.6292
Royalty	0.549	0.5243	0.5367
Sport	0.5156	0.498	0.5068
Warfare	0.606	0.4893	0.5477

than a single source (such as only text or image). Figure 6.10 is similar to Figure 6.9 but in case of uni-modal retrieval task using traditional SOM implementation. Sorted precision values' curve for the retrieval of matched images using an image query by utilizing 12-D ZM and 128-D SIFT features is presented in Figure 6.10a, however, Figure 6.10b correspondingly demonstrates the similar curve for the retrieval of matched text using a textual query by utilizing 14-D LDA and 10-D LDA features.

6.1.7 Discussion

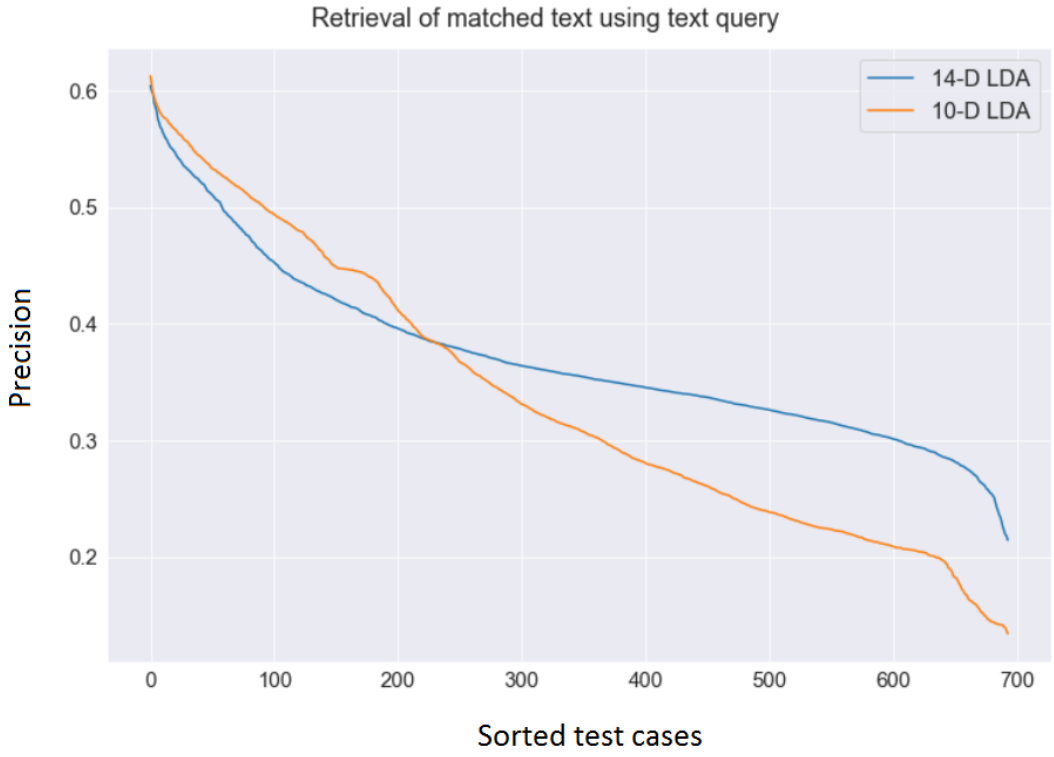
The proposed approach shows better performance than all the compared methods including baselines and other state-of-the-art methods because of the following reasons:

(a) I2T operation

(b) T2I operation

Figure 6.9: Curves depicting the sorted precision values for test queries (693 in our case)

(a) I2I operation



(b) T2T operation

Figure 6.10: Precision-scope curves for uni-modal retrieval

Table 6.5: Comparison of MAP scores obtained using traditional SOM and HSOM

Methodology	**Task**	**Features**	**MAP Score**
Traditional SOM	I2I	128-d SIFT	0.4028
		12-d ZM	0.367
	T2T	10-d LDA	0.3347
		14-d LDA	0.3716
Hybrid SOM	I2T	128-d SIFT, 10-d LDA	0.5872
		12-d ZM, 14-d LDA	**0.6461**
	T2I	128-d SIFT, 10-d LDA	0.4744
		12-d ZM, 14-d LDA	**0.5228**

1. ZM have been utilized for visual feature representation which are least redundant, noise resilient, and rotation, scale, and translation invariant. They capture the global image features and also effectively describe the shape characteristics of an object in an image [171].

2. LDA features that are used for text representation provide well-defined inference procedures for even unseen documents [172]. They reveal the inter and intra document statistical structure. An appropriate value for the number of topics in LDA model implementation has been decided based upon the perplexity analysis.

3. SOM provides potent clustering of images and text as it emulates the working of neurons inside the human brain [190]. Moreover, SOM has shown its effectiveness recently in various application areas such as speech recognition [235], mental stress detection [236], coronary heart disease diagnosis [237], and for extracting features as an add-on for better network intrusion detection [238].

4. SOM helps in easy interpretation and understanding of data [190]. Similarities in data can be easily observed and visualized using SOM. It has the ability to cluster even large or complex datasets [191].

5. The Hebbian network which is used for integration of image and text SOM to make coordination between the modalities works on the principle of Hebb's rule which is also inspired by biological systems [193]. The goal of the proposed technique is to construct a system which, on giving an image query, can find the matched images and provides a suitable annotation to it like humans.

The proposed technique outperforms the compared *deep learning* based methods due to the following reasons:

1. As per [239], numerous processes in deep convolutional neural network are nowhere near to the ones that happen inside brain. For instance, the training process in deep neural network is based on backpropagation and stochastic gradient descent optimization, however, neuroscience suggests that biological brain does not have these kind of processes. Instead, learning approaches which are based on Hebb's learning rule or Spike-timing-dependent plasticity appear to be more reasonable.

2. A huge dataset (sometimes in millions) is required for deep learning techniques to work well [240] which is not present in this study, otherwise it may result in overfitting during model training and thus may not perform well on the test data [241].

3. Typically, a deep learning method (such as CNN) cannot directly perform better than the machine learning methods. Its performance highly depends on the design which includes input window size, layer depth, and training strategies [242].

4. Training the pre-trained model again from scratch might not be feasible as it requires the understanding of a large number of model parameters and the modifications in layers which is again computationally expensive as well [243].

5. The selection of appropriate feature extractors for the modalities also comprises a considerable part of the whole algorithm. Representation of the modalities must be done suitably to enhance the overall system performance [244]. That is why Zernike moments and Latent Dirichlet allocation features have been utilized with appropriate parameter values so that the modalities can be represented in the best possible way.

6.2 Case study on primary endoscopy dataset

This section includes the data analysis of the endoscopy data, results obtained from clustering of endoscopic images using SOM, and application of the proposed *HSOM* approach to this data using both the Hebb and Oja rule. Here, Wavelet transforms (WT) and Zernike moments (ZM) have been utilized for image representation for clustering implementation. VGG16 deep features are extracted from the images for the cross-modal retrieval task, and Term Frequency Inverse

Document Frequency (TFIDF) features are used for collateral text representation after pre-processing. Separate Self Organizing Maps (SOM) have been trained using the generated image and text features, which are then associated using Oja links. Then the network (created through Oja links) between trained image and text SOM is trained for the total number of input instances to create the final cross-modal retrieval system capable of retrieving the same as well as different data than the query data. Diverse quantitative comparisons have been performed on the gastrointestinal data, utilizing different combinations of image and text features and the different learning rules such as Oja and Hebb for modalities' integration network training.

6.2.1 Dataset

The dataset consists of 300 real gastrointestinal images accompanied with collateral text obtained from a known gastroenterologist. The dataset has a ratio of 180:120 for healthy and sick cases. Figure 6.11 shows an image-text pair instance from the dataset. All the images are of size 256×256 and the data incorporates four diverse categories of in-vivo gastral images: Upper (normal and bleeding) and Lower (normal and bleeding) as shown in Figure 6.12. Upper gastrointestinal tract includes *esophagus and stomach*, lower includes *small bowel and colon*. These areas have been further divided into normal (180) or bleeding (120) instances. In upper GI tract, normal = 42, bleeding = 54; in lower GI tract, normal = 138 and bleeding = 66 instances. Table 6.6 provides the information regarding the endoscopic images in the dataset.

Image	Collateral text
	PillCam Small Bowel gluten refractory celiac disease

Figure 6.11: Sample dataset instance: a gastral image and collateral text

6.2.2 Data distribution analysis

The data distributions of healthy and sick image vectors have been examined from various aspects (to analyze an appropriate learning model), which are as follows:

- Relative red intensity in healthy/sick images.

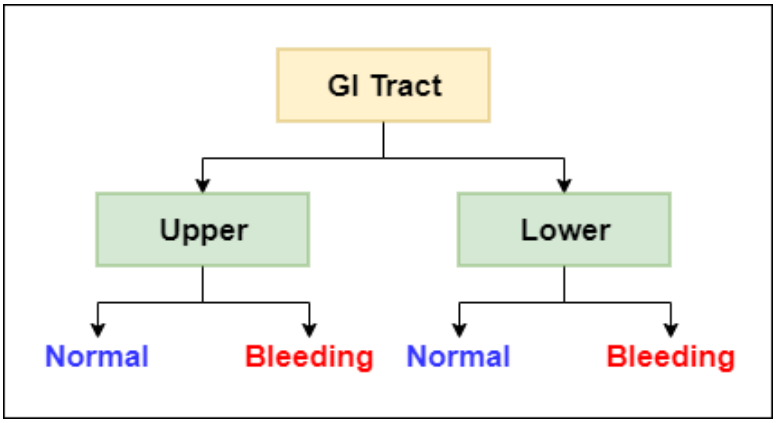


Figure 6.12: Dataset of 300 images with collateral text has been divided into upper (esophagus, gastric) and lower (small bowel, colon) sections. Upper: normal = 42, bleeding = 54; Lower: normal = 138 and bleeding = 66

Table 6.6: Description of images in dataset

Category	**Healthy**	**Sick**
Image ratio	180	120
Size	256×256	256×256
Redness	Overall	Spots or saturation

- Distribution of RGB intensities in all images.
- Thresholding to crop the red color (for example, $R > 100, G < 60, B < 50$).

In Figure 6.13 the average red color in the sick and healthy classes has been sorted and plotted. Although all gastral images are reddish brown in color, the sick images are more saturated with redness. The intensity plots of all the images have been illustrated in Figure 6.14: (a) shows that the left half has more dispersion, especially in red color and R values are relatively higher. The other three plots (b-d) show the R versus G, R versus B, and B versus G plots. There is tremendous overlap, so a simple linear regression may not be sufficient for the bleeding analysis. Thus, there is a need for a non-linear learning system (such as SOM). The red color segmentation has been experimented with using MATLAB to further analyze the problem complexity, which is illustrated in Figure 6.15 . The threshold values $R > 100, G < 60, B < 50$ have been chosen for best human eye subjective red color cognition. Again, it seems quite difficult to distinguish the healthy red versus the sick red spots for confounding cases. The middle two images are healthy in these four images, and the left/right extremes are bleeding cases. Therefore, simple thresholding is also insufficient to spot the bleeding even with various threshold values.

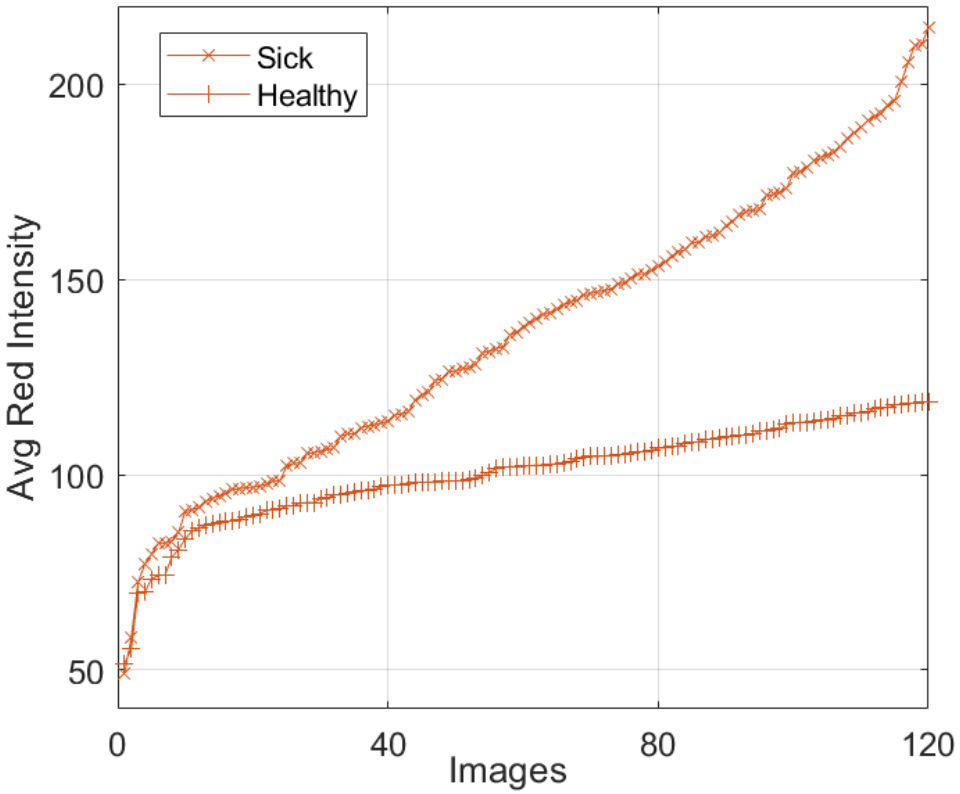


Figure 6.13: Sorted average red pixel intensities for normal and abnormal images. The upper line is redness in sick images which is relatively higher as compared to healthy images.

6.2.3 ZM extraction and SOM application

For each L, U, and V component, 12 ZM have been calculated, making a total of 36 ZMs to extract the luminance and color attributes as shown in Figure 6.16. Zernike moments are rotation and noise invariant as studied in [186], which can be seen in the figures (6.17,6.18). Furthermore, feature transformation techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are applied to these 300×36 image vectors as shown in Figure 6.19. Both of these transformations are used for dimension reduction, but PCA focuses on maximizing the variance among mutually orthogonal transformed dimensions, and LDA focuses on the separability of the data concerning the labels [245]. Linear separability in the case of PCA has been found to be 74% and 84% in case of LDA. Therefore, linear separability becomes possible after extracting ZM on LUV image components. Self-Organizing Maps have been involved further to analyze the accuracy with the hope of improvement.

Figure 6.20 shows the results when the model was trained using only healthy (180) points and tested for 120 sick images. It can be observed that there is some overlap of the tested sick images with healthy images due to the obscure nature

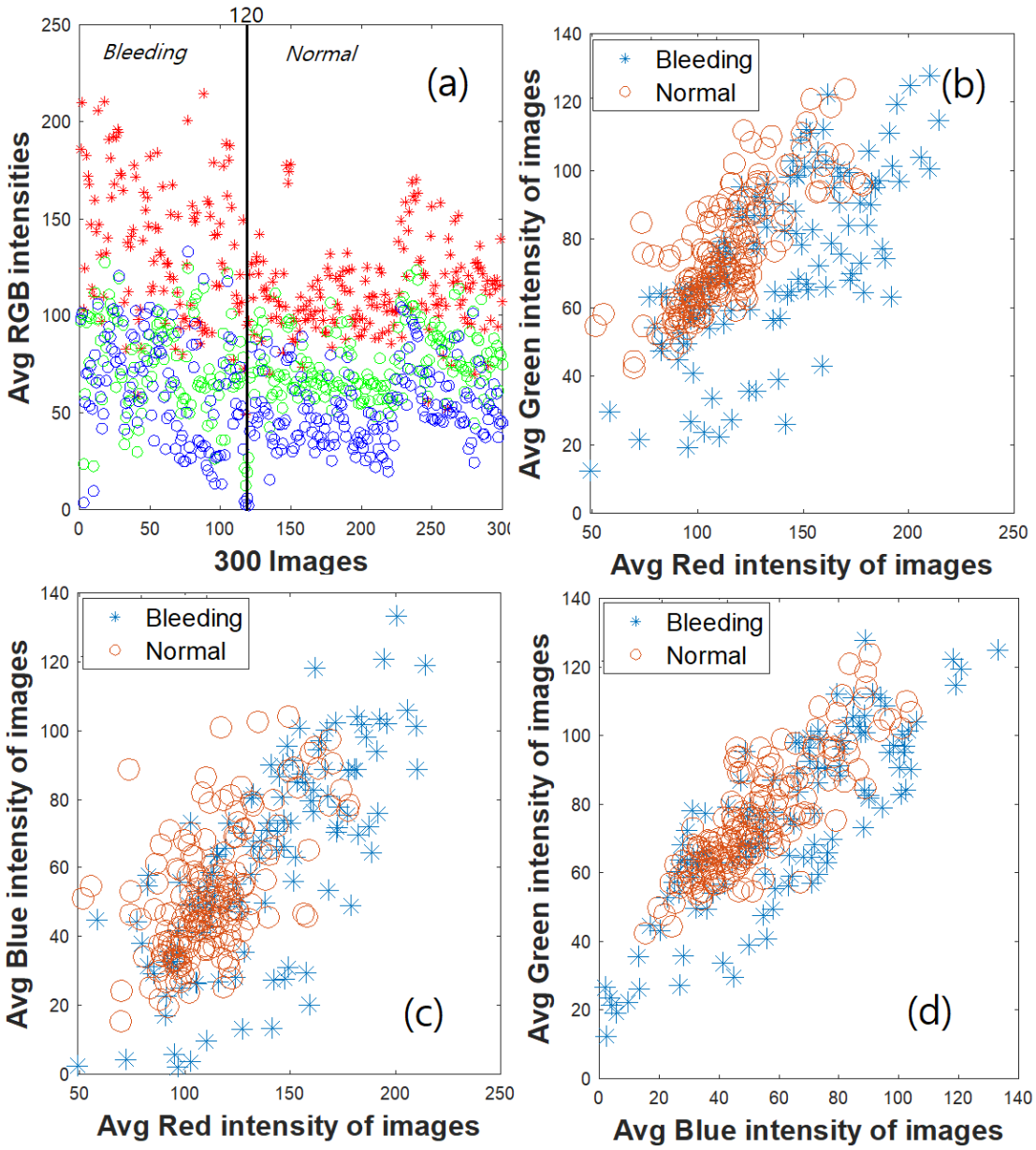


Figure 6.14: (a) RGB intensity of 300 images (180 : 120 for healthy:sick cases) (b) Red vs Green average intensities (c) Red vs Blue avg. intensities (d) Blue vs Green intensities plotted for all the images. It is clear from (b-d) that none of the color intensities are easily separable for healthy and sick cases.

of the images. But still, there is a complementary saturation between these two images showing that sick images have different data distribution on a broader scale.

Wavelet transforms applied to all the images before extracting their Zernike moments (ZM) for noise removal. In Table 6.7, the results of accuracy for ZM versus WT+ZM on 300×36 image vectors have been compared. Table 6.8 shows the difference between WT+ZM and ZM accuracy is positive on the average of 5 trials,

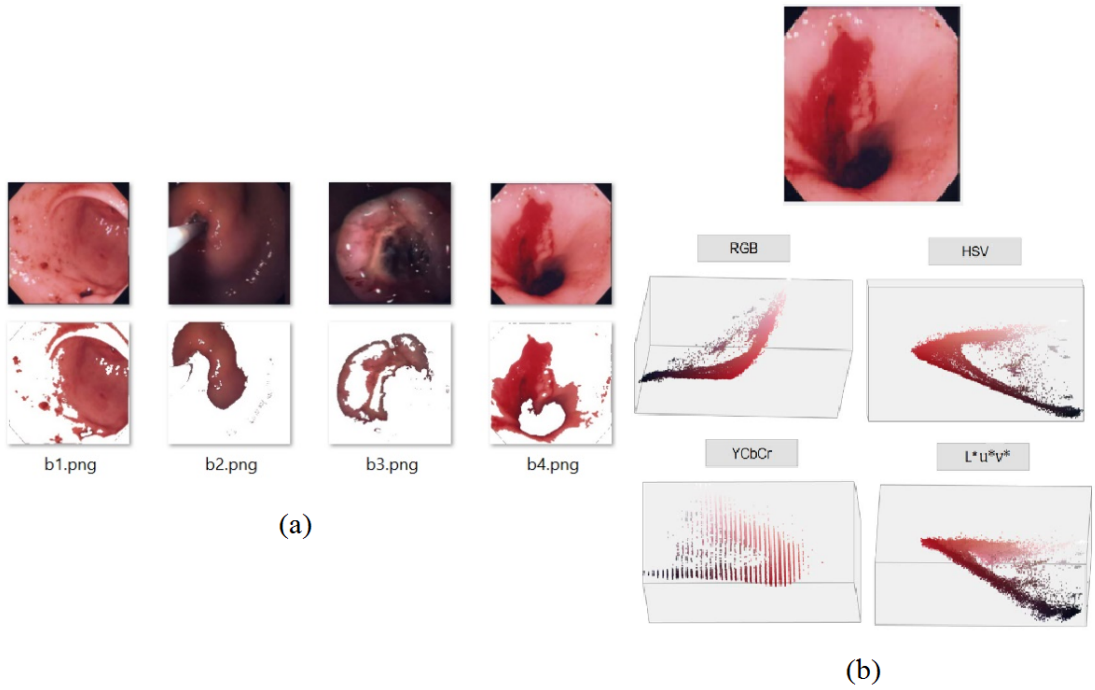


Figure 6.15: (a) Red color cropping using subjective threshold RGB values (using MATLAB) $R > 100, G < 60, B < 50$. Hence, in order to classify the images, robust features are required to extract such as ZM. Firstly, RGB values have been converted to LUV color representation to separate the luminance component from the color composition as shown in Figure (b).

	1	2	3	4	5	6	...	34	35	36
n1	512.4	165.4	204.8	6.6	144.3	36.6		34.5	17.8	0.2
n2	533.6	165.3	243.5	4.6	157.9	31.0		10.0	5.7	0.9
n3	457.5	141.7	217.3	1.3	143.5	19.6		31.7	11.1	1.1
n4	444.7	139.3	204.2	1.2	136.0	21.8		25.7	12.3	0.0
n5	520.9	157.2	239.8	11.5	154.9	30.0		30.1	16.6	0.6
n6	518.1	158.5	252.4	3.0	162.1	21.2		39.3	18.4	0.3
n7	476.4	150.6	218.2	5.1	146.0	22.5		30.1	8.3	0.1
n8	555.1	170.9	263.1	0.9	171.3	25.1		19.5	8.1	0.3
n9	470.4	143.6	215.9	3.8	131.2	28.2		22.5	9.8	1.1

Figure 6.16: Snapshot of ZM for 9 healthy images. Rows corresponds to the image vectors and 36 columns are the Zernike moments of image. n means normal/healthy.

confirming the advantage of WT application before ZM extraction. Figure 6.21 is the visual illustration of Table 6.8. Finally, Table 6.9 presents the confusion matrix obtained after experiments and the best accuracy obtained using WT+ZM and Kohonen self-organized feature maps has been found to be approximately 88.3%. In the table, P_Healthy, P_Sick and A_Healthy, A_Sick are predicted and actual values of healthy and sick classes respectively. Table 6.10 shows the comparison

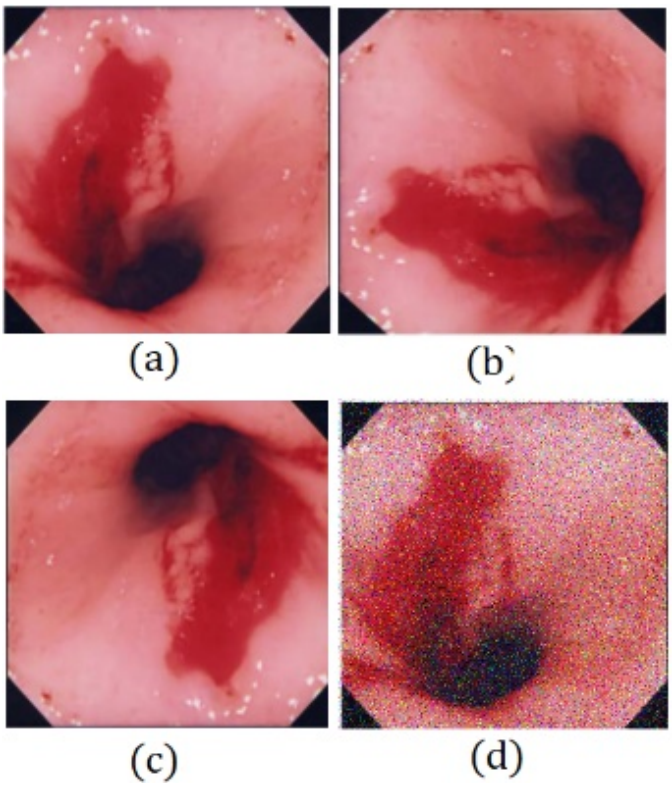


Figure 6.17: (a)-(d) are original image, rotations of 90 and 180 degrees, Gaussian noise introduced.

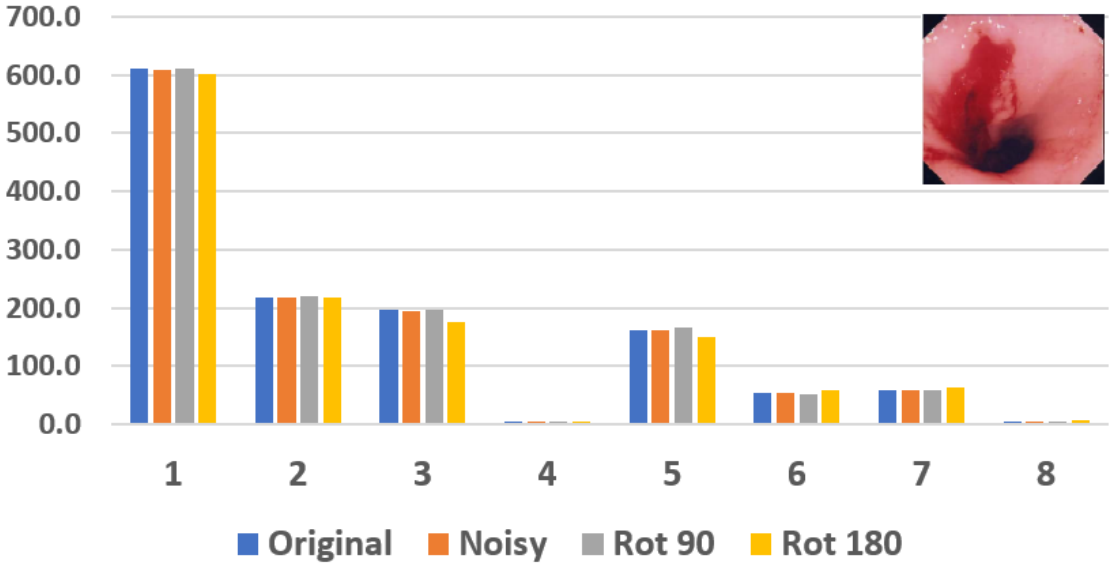


Figure 6.18: ZM for all images in Figure 6.17 are nearly same. Only 8 out of 36 ZM have been shown to simplify the demonstration. Slight differences are due to 3 types of errors involved in ZM calculation as studied in [184].

of the other approaches with the proposed technique based on obtained accuracy. The comparison is performed with both traditional methods such as PCA and LDA (used for linear separability of data), and contemporary methods such as deep learning based approaches.

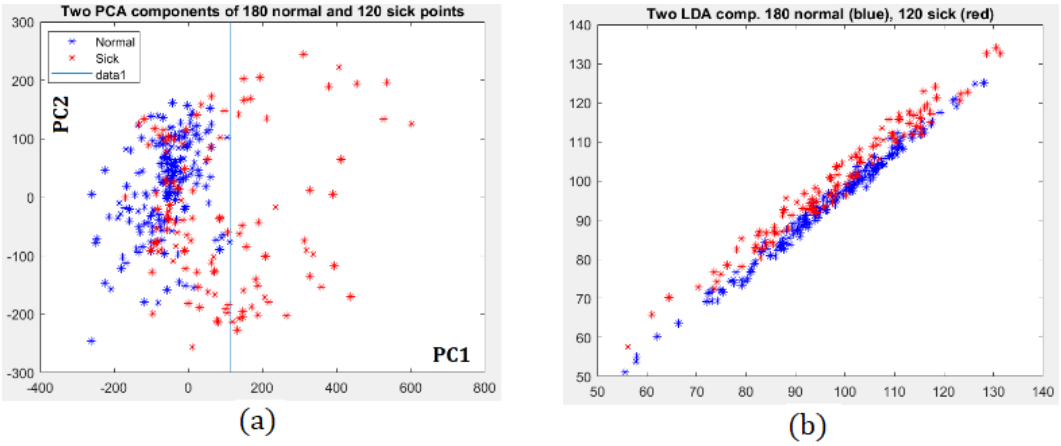


Figure 6.19: PCA and LDA draws of (300) image vectors. PCA yields 74% and LDA yields 84% linear separability of 120 bleeding and 180 normal image vectors. *Blue* for healthy and *Red* for sick.

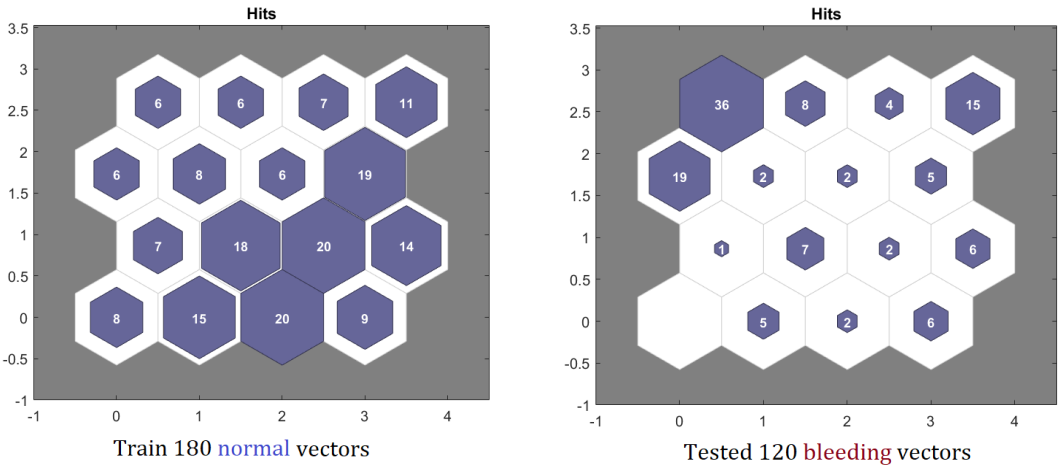


Figure 6.20: SOM maps obtained upon training with 180 normal points and testing with 120 bleeding images using MATLAB. It can be observed that the mapping is different and thus distribution of 180×36 and 120×36 image vectors are different.

Single SOM illustration

We begin by looking at the performance of a single SOM to deal with the categorization of images. Figure 6.22 is a collection of 10 healthy and 10 sick gastral images along with their captions. A simple SOM has been executed for individual image and text clustering for the 4×4 map as shown in Figure 6.23 and (6.24) respectively. The image category has subtle visual features that are shared among each member: color, texture, shape, and edges. Hence, the clustering of gastral images together is no accident as depicted in Figure 6.23.

The testing for both Single and Hybrid SOM has been explained using 300 instances (image + collateral text) with 90 : 10 for training and testing ratios in the following sections.

Table 6.7: Accuracy given by ZM versus WT+ZM on 300 × 36 images. TR = training data, TS = testing data, VAL = validation data, and AVG = average.

	ZM				**WT+ZM**			
Trials	**TR**	**VAL**	**TS**	**All**	**TR**	**VAL**	**TS**	**All**
1	82.4	78.3	71.7	78.7	83.8	86.7	75.6	83
2	79.5	84.4	77.2	80	81.4	86.7	77.8	81.7
3	77.6	82.2	77.8	78.3	82.4	81	76.9	81
4	75.2	80.2	77.8	76.7	82.4	80	80	81.7
5	80.5	82.2	80.2	81	83.3	88.9	79.6	83
AVG	79	81.5	76.9	78.9	82.9	84.5	79.8	82.4

Table 6.8: Difference in the accuracy values of after and before WT while extracting ZM. Positive values in the last row signifies the benefit of WT application prior to ZM.

	Accuracy of WTZM - ZM			
Trials	**TR**	**VAL**	**TS**	**All**
1	1.4	8.4	3.9	4.3
2	1.9	2.3	0.6	1.7
3	4.8	-1.2	-0.9	2.7
4	7.2	-0.2	2.2	5
5	2.8	6.7	-0.6	2
AVG	3.8	3	2.9	3.4

Table 6.9: Average test results for 5-trials with the proposed method on 300 image vectors. P_Healthy, P_Sick and A_Healthy, A_Sick are predicted and actual values of healthy and sick classes respectively.

	A_Healthy	**A_Sick**	**Total**
P_Healthy	161	16	177
P_Sick	19	104	123
Total	180	120	300
Accuracy	89.4%	86.7%	88.3%

Testing single SOM

For an image to image retrieval, the overall recognition accuracy for Single SOM has been found to be 77% as shown in Table 6.11. The results shown are the

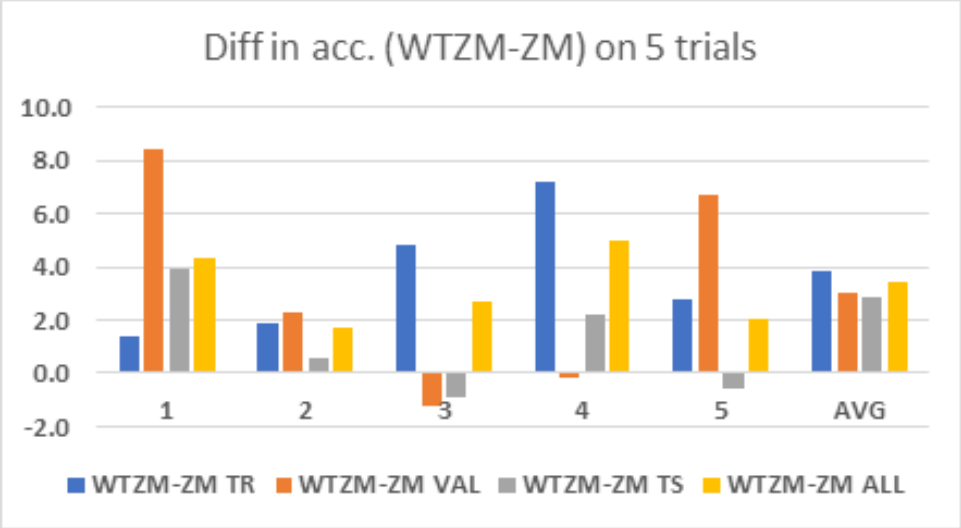


Figure 6.21: Difference in the accuracy of results by using WT+ZM versus only ZM as in Table 6.8.

Table 6.10: Comparative analysis of techniques on the underlying dataset.

Sr.	**Reference**	**Technique**	**Accuracy**
1	[214]	Principal component analysis	74%
2	[203]	Agglomerative clustering and CNN	80.5%
3	[204]	Nonlinear subspace clustering	83.3%
4	[245]	Linear discriminant analysis	84%
5	[207]	Robust learning for unsupervised clustering	85.7%
6	[208]	Deep density-based image clustering	86.8%
7	[205]	Hierarchical clustering using 1D random projections	87.1%
8	[206]	Deep adaptive image clustering	87.6%
9	[209]	Discriminative pseudo supervision clustering	87.9%
10	**Proposed**	**WT+ZM on LUV**	**88.3%**

average of 5 runs. The keyword categorizing SOM performs a relatively simple task in that it matches keywords with keywords, and the accuracy of the retrieval for 30 keyword vectors presented 5 times was almost 100%. It should be noted here that choice of images and the assignment of categories by dataset indexers varied considerably in detail and perhaps in accuracy, suggested by the variation in the number of keywords, and some labels described visual features like ulcer,

	SICK	Caption (b_i)	NORMAL	Caption (n_i)
1		Ulcer gastric antrum		SB AmpullaofVater
2		Hemostasis ulcervessel		SB normal mucosa
3		Ulcer posthemostasis		SB celiac disease
4		Bleeding anal esophagogastric		SB celiac disease
5		Hemostasis bleeding		SB celiac disease
6		Tear posthemostasis		Esophageal varices ESO
7		Bleeding anal esophagogastric junction		Esophageal varices ESO
8		Hemostasis bleeding		Esophageal varices ESO
9		Tear posthemostasis Ulcer gastric antrum		Esophageal varices ESO
10		Ulcer bleeding gastric antrum		esophageal varices

Figure 6.22: A small example of 20 images and 20 captions. Here b_i and n_i means the corresponding captions of bleeding and normal examples.

Angioectasia and so on.

6.2.4 Testing hybrid SOM exploiting Hebb rule and ZM

The hybrid system is trained in a way as mentioned in *Chapter 3* and *Algorithm 3*. The trained system is capable of retrieving both the same as well as different modality than the query modality. However, the main focus of a cross-modal retrieval is to analyze the accuracy of a system based on the efficiency of retrieving different modality than the query modality. So, the two tasks that are being automated here are: (1) *Auto-annotation*, and (2) *Auto-retrieval*

1. *Auto-annotation*: The trained hybrid system, comprising an image categorization system and a keyword categorization system connected via a Hebbian link, is presented with an unseen image feature vector. The image SOM

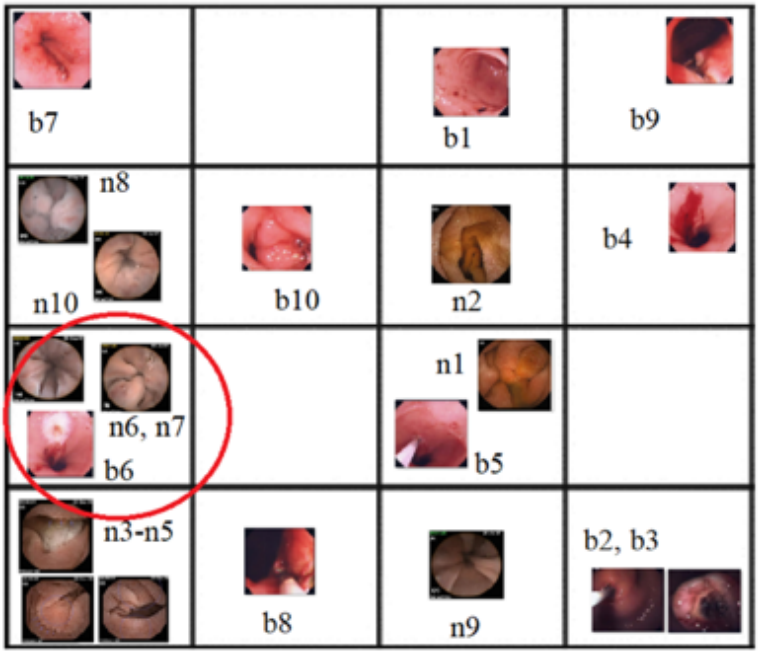


Figure 6.23: Clusters created by the image features (ZM) in the SOM on 20 images (Figure 6.22) for 4×4 grid of SOM. The circle shows overlapping node with normal and bleeding images.

	n6-n9: esophageal varices	**b2: hemostasis ulcervessel b5, b8: hemostasis bleeding**	
n3,n4,n5: SB celiac disease	n10: esophageal varices	n: SB ampullaofvater n2: SB mucosa	**b4, b7: bleeding anal esophagogastric**
			b3: ulcer posthemostasis b6: posthemostasis tear b9: ulcer gastric antrum posthemostasis tear
	b1: ulcer gastric antrum b10: ulcer gastric antrum bleeding		

Figure 6.24: Clusters created by the text features in the 4×4 SOM for the sample data shown in Figure 6.22. n and b means normal and bleeding instances respectively. The keywords next to n and b are the keywords corresponding to that particular instance.

component is activated by the unseen image vector, and the system finds the best matching units (BMUs). The active nodes then activate nodes in the collateral keyword SOM via the Hebbian links. The hybrid system determines the BMUs. The auto-annotation process is shown in Figure 6.25. It can be seen that the system retrieves correct terms, additional keywords, and spurious keywords. The spurious keywords include those words in the training set that may belong to an image either in a different category or a different image within the same category but with similar visual features.

Table 6.11: Accuracy scores of different retrieval operations using single SOM and hybrid SOM on endoscopy data

Retrieval operation	**Accuracy (round off)**
I2I	77%
I2T	83%
T2I	85%

For instance, when a *celiac disease* image was presented (Sr. 2 of figure 6.25), the system retrieved 3 different $n3 - n5$ images. The activated nodes in the keyword SOM contained 4 terms that were collateral to the presented unseen image ("*pillcam small-bowel celiac disease*"). As an output, it got 5 total keywords which includes one spurious keyword "*pillcam*". The average accuracy obtained for auto-annotation task is 83% (Table 6.11).

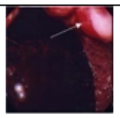
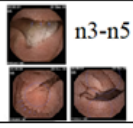
Sr.	**Image**		**Collateral Keywords**		
	Query	**Retrieved**	**Correct Keywords**		**Spurious Keywords**
			Matched Keywords	**All Keywords**	
1		 b10	Ulcer gastric antrum bleeding	Upper gastrointestinal Ulcer bleeding gastric antrum	
2		 n3-n5	Pillcam small-bowel celiac disease	Small-bowel celiac disease pillcam	pillcam
3		 n2	Small-bowel pillcam mucosa	normal mucosa Small-bowel pillcam	pillcam

Figure 6.25: Auto-annotation testing results

2. *Auto-retrieval*: In this case, the hybrid system is given an unseen query keyword vector. The system matches the keywords with the pre-stored nodes in a keyword categorizing SOM. The identification of the place in which query-related keywords are concentrated then allows the search for collateral images whose features have been 'learned' by a SOM. We present a selection of auto-retrieval examples in Figure 6.26. Given a query, the retrieved keywords are divided into three parts: matched words, all related keywords, and spurious words. The average accuracy is 85% for image retrieval task (Table 6.11).

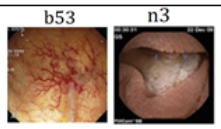
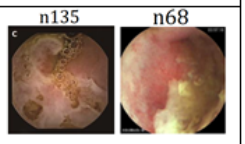
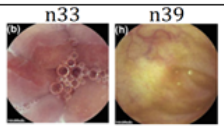
Query Keywords	**Retrieved Keywords**			***Example* Collateral Images**
	Matched	**All**	**Spurious**	
Small bowel polyp	Bowel polyp	Ileum hyperplasia celiac SB bowel polyp angiectasias omom	omom	
Crohnsdisease Serpiginous ulcer	Crohnsdisease ulcer	Serpiginous ulcer Crohnsdisease ulcer Bleeding gastric diverticuli	gastric	
normal lesion Gastric erosion	Normal lesion	Duodenal polyp Stricture lesion GI posthemostasis normal	GI	

Figure 6.26: Testing results for auto-retrieval. Image $b53$ means bleeding image number 53 and $n3$ means normal image number 3. Only 2 images have been shown to simplify the illustration.

Results reported for the hybrid SOM method in Table 6.12 represent the average values of class-wise accuracy over the test images for different retrieval tasks. The bold accuracy values depict the best value of accuracy on that particular data class and the retrieval task. The chart in Figure 6.27 represents the category-wise accuracy scores for various retrieval tasks. It can be observed from the chart that the best accuracy is obtained in case of both *Upper GI - Bleeding* and *Lower GI - Bleeding* categories for *T2I* and *I2T* retrieval tasks respectively.

Table 6.12: Comparison of the accuracy for the 4 categories of images (I) and collateral text (T). Bold numbers indicate the best value of accuracy.

Category	I2I	T2I	I2T	Avg	Best Task
Upper GI - Normal	77%	80%	**87%**	81%	I2T
Upper GI - Bleeding	73%	**93%**	80%	82%	T2I
Lower GI - Normal	77%	77%	**80%**	78%	I2T
Lower GI - Bleeding	80%	83%	**93%**	85%	I2T

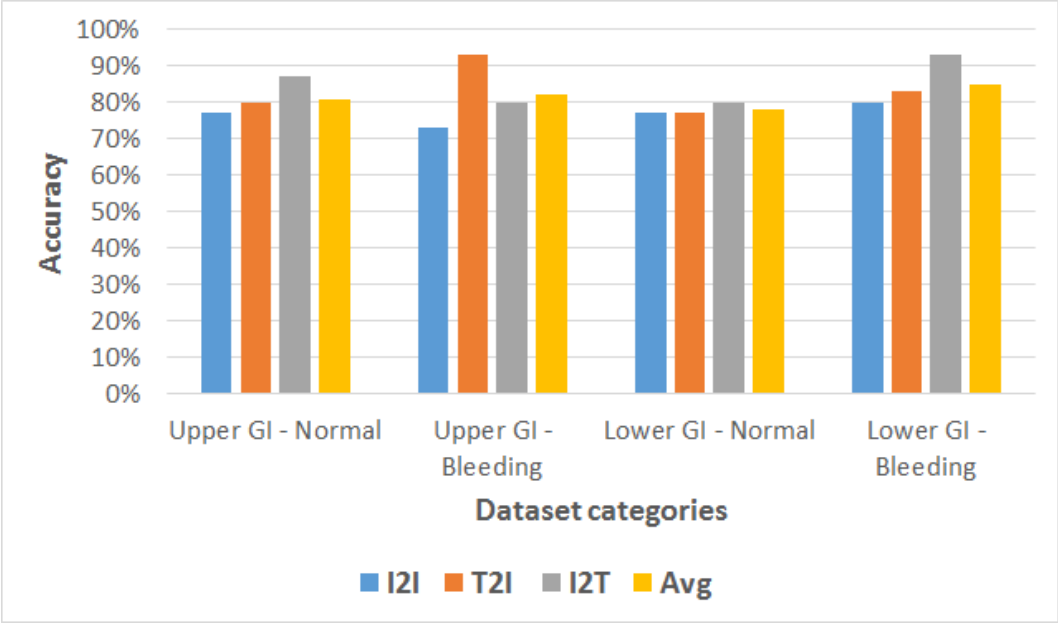


Figure 6.27: Class-wise performance chart on different retrieval tasks based on Accuracy scores

6.2.5 Testing hybrid SOM exploiting Oja rule and deep features

The detailed description and steps required for the implementation of this approach (dubbed $HSOM\_OJA$) are given in *chapter (5)* and *Algorithm 5*. Here, the VGG16 deep features are extracted from the endoscopy images, and TFIDF features represent collateral text. Then the separate SOMs are trained using these visual and textual features. The two SOMs are integrated using an Oja network or improved Hebbian network following the Oja learning rule to make a final cross-modal system that can be utilized for gastrointestinal image annotation and retrieval. The implementation details are given in the following sections. The final results obtained using this $HSOM\_OJA$ approach are compared with the results obtained using HSOM exploiting the Hebb rule, ZM, and LDA (dubbed $HSOM\_HEBB$).

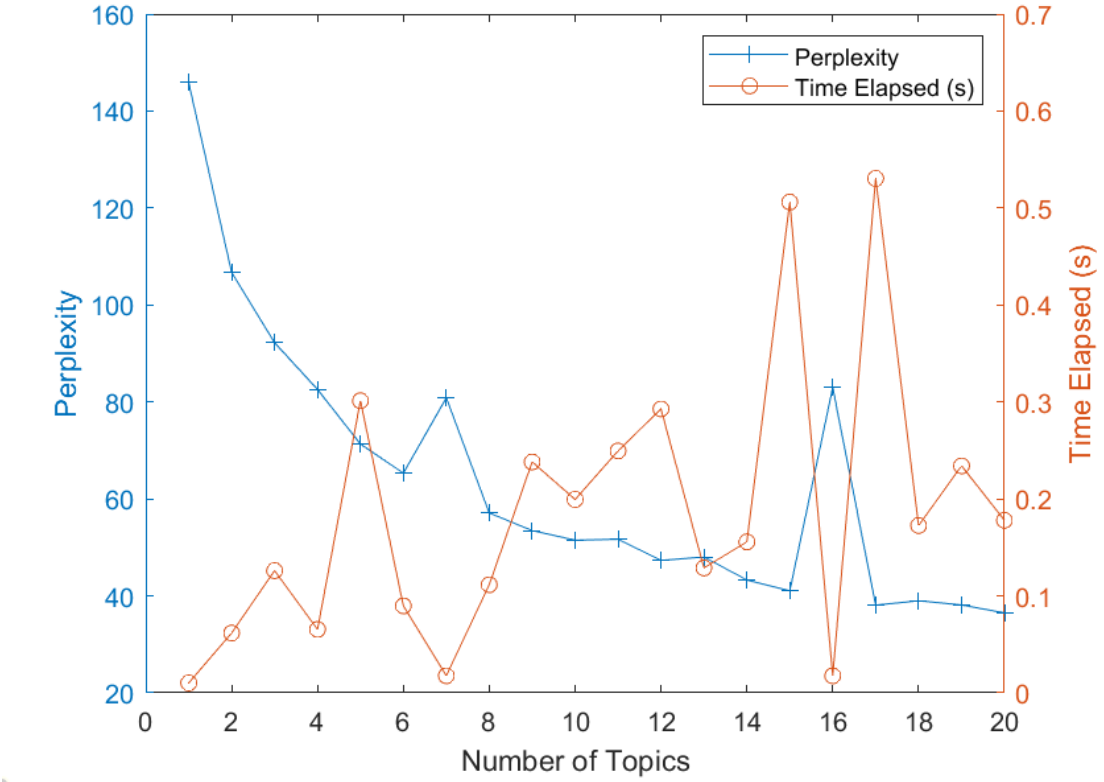


Figure 6.28: Perplexity and time analysis to determine an appropriate number of topics for the LDA model in endoscopy data.

Parameter settings

The parameter values of the different methods utilized in the implementation are chosen to enhance the system's overall performance. Order 5 has been chosen for the ZM features used in $HSOM\_HEBB$, generating a 12d image vector. Perplexity and time analysis has been performed to determine the appropriate value for the total topics created in the LDA model for textual representation. This analysis has been represented in Figure 6.28 in the form of a plot. *Perplexity* is the statistical measure of how well a probability model predicted a sample. The number of topics needs to be selected such that the perplexity value is minimized. The lesser the perplexity value, the better the selection of the number of topics. However, the convergence time of the LDA model also needs to be considered along with perplexity because, with an increase in the number of topics, LDA may require more time to converge. So, to critically analyze this trade-off, both the perplexity score and the elapsed time are plotted simultaneously against the number of topics as depicted in Figure 6.28. As per the figure, the best choice for the total number of topics is 14, which has been selected for the LDA model training, and it generates the 14d text vector corresponding to each textual instance. For extraction of deep image features in the $HSOM\_OJA$ technique, a pre-trained VGG16

convolution neural network has been used with default parameters. The features have been extracted from the $fc8$ fully connected layer of the model generating a 1000d feature vector for image representation.

Model training

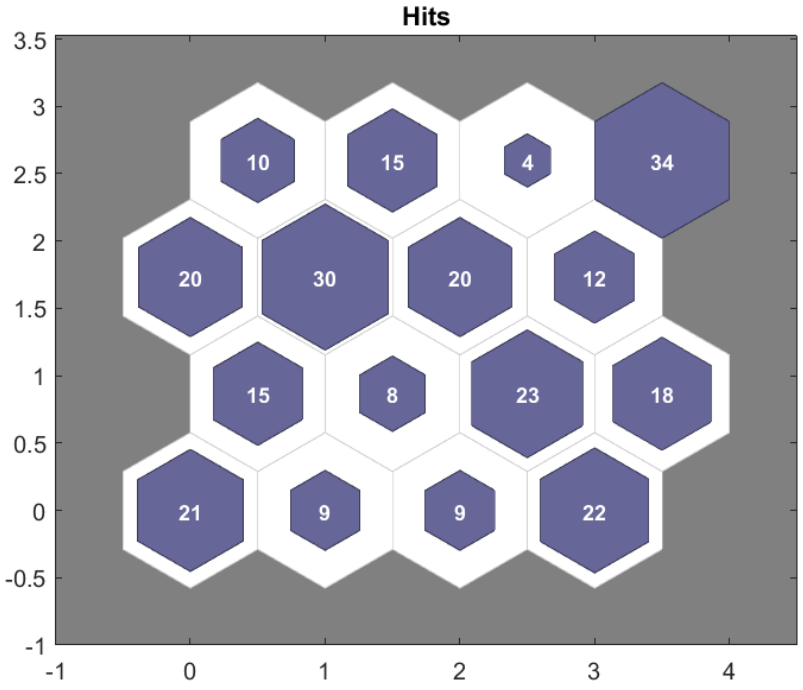
Figure 6.29 shows an example of the train data distribution after independent training of image and text SOM. The hexagon represents the neuron or a node in the SOM, and the number written inside each node depicts the total number of train instances clustered in that particular SOM node. These SOMs are trained using 1000d VGG16 visual features and 14d LDA (Latent Dirichlet Allocation) textual features. The corresponding SOM figures depicting neighbor distances are demonstrated in Figure 6.30. Red lines connect SOM nodes (depicted as blue hexagons) to the neighbors. The darker the shade of the color in the red line section, the more is the distance between the adjacent nodes. The prominent tuning parameters chosen for image-text SOM training (after several experiments) in MATLAB are given in Table 6.13. The learning rate for the Oja network has been chosen by training multiple times at different values $\{0.1, 0.01, 0.001\}$ and analyzing the average image-text query MAP score. After experimentation, 0.001 has been found to be an appropriate learning rate for the endoscopy dataset.

Table 6.13: SOM parameters chosen for experimentation (in MATLAB) using endoscopy data

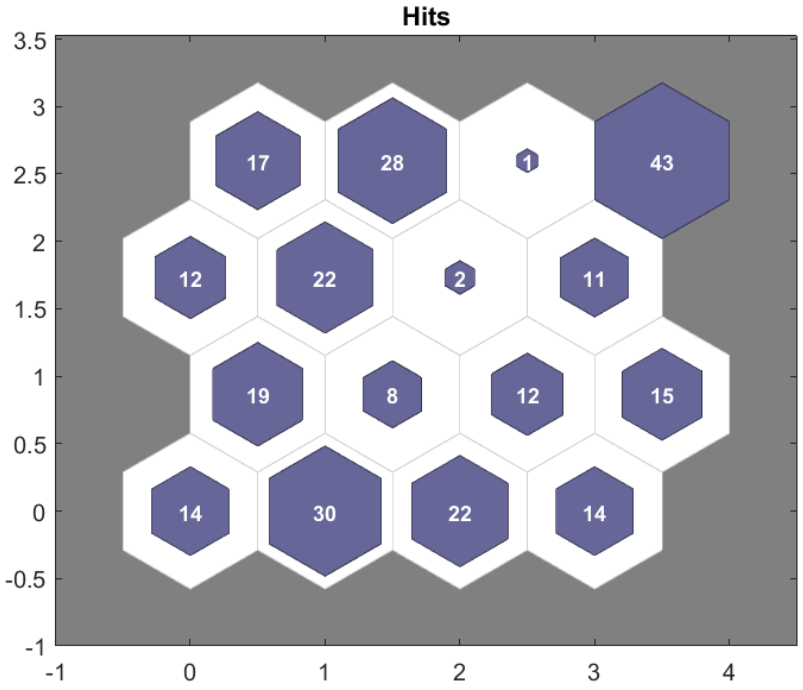
Parameter	**Value**	**Description**
dimensions	[4 4] (4×4)	Row vector of SOM dimension sizes
coverSteps	100	No. of training steps for initial covering of input space
initNeighbor	3	Initial neighborhood size
topologyFcn	hextop	Layer topology function
distanceFcn	linkdist	Neuron distance function

Results

Table 6.14 shows the comparative analysis of different techniques based on the MAP score performance metric. Diverse combinations of image and text features have been utilized along with the two SOM association network weight updation rule. The table consists of five columns. *Rule* column represents the weight updation rule followed for training the image and text SOM integration network,



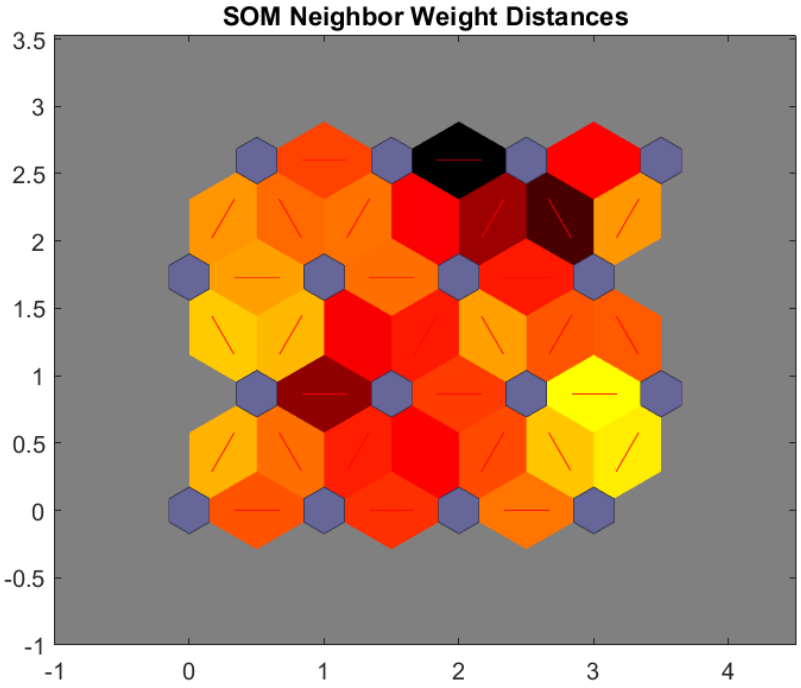
(a) Trained image SOM



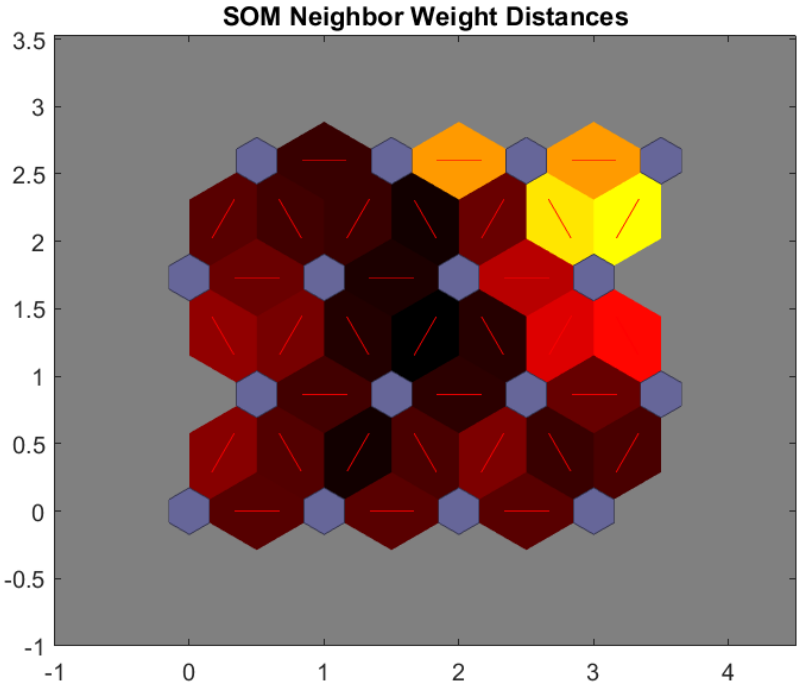
(b) Trained text SOM

Figure 6.29: Input train data distribution after individual SOM training for endoscopy data

such as the Hebbian and Oja learning rule. The columns *Image features* and *Text features* represent the various image and text representation methods used for



(a) Trained image SOM



(b) Trained text SOM

Figure 6.30: Neighbor Distances among respective SOM nodes after training for endoscopy dataset. Darker shade denotes larger distance.

the experiments, which are 12d ZM, 1000d VGG16, 14d LDA, and 323d TFIDF features. MAP_I2T and MAP_T2I columns represent the MAP score values for

Table 6.14: MAP score comparison of different combinations of methods on endoscopy dataset.

Rule	Image features	Text features	MAP_I2T	MAP_T2I
Hebb	ZM	LDA	0.6435	0.5451
		TFIDF	0.6694	0.5533
	VGG16	LDA	0.5824	0.4304
		TFIDF	0.5811	0.5591
Oja	ZM	LDA	0.6574	0.5297
		TFIDF	0.5472	**0.5767**
	VGG16	LDA	**0.6897**	0.3498
		TFIDF	0.6073	0.5376

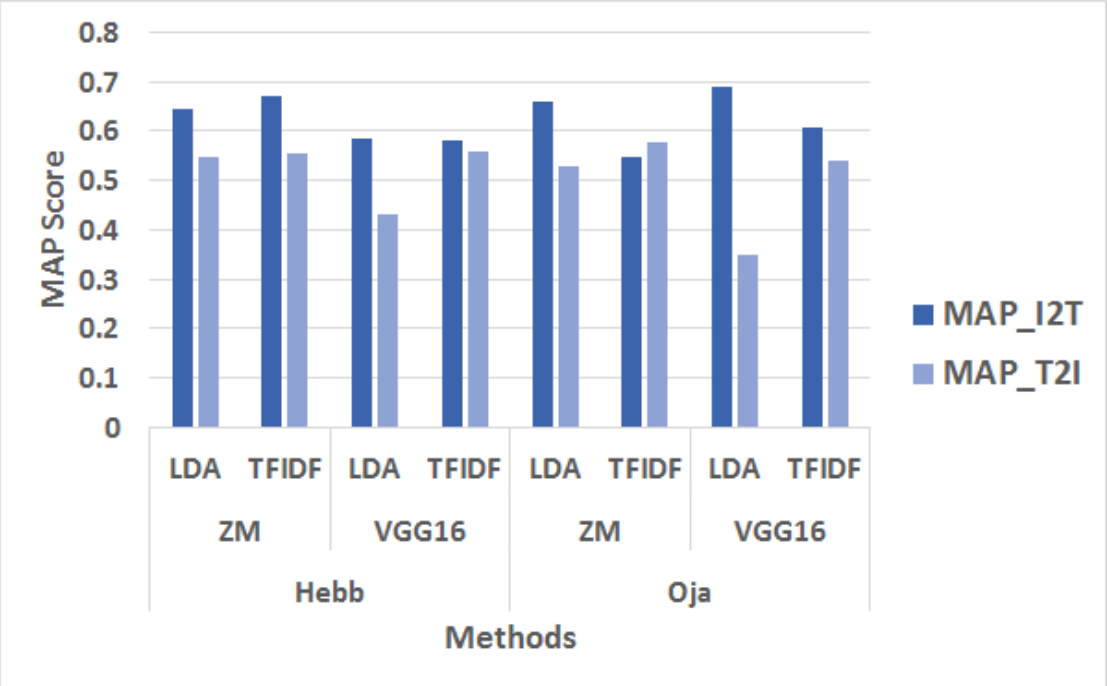


Figure 6.31: Performance analysis of different methods based on MAP score.

image-to-text retrieval (image annotation) and text-to-image retrieval (image retrieval), respectively. It can be observed that the Oja network training, along with VGG16 image features and LDA text features, is showing the best performance in image-to-text (I2T) retrieval operation, however, it is showing the worst performance for the T2I task. So, it can be deduced that the combination of VGG16 and LDA along with the Oja rule is not good for the overall system performance. Oja rule with ZM image representation and TFIDF text representation is giving the best performance for text to image retrieval (T2I) task. The average (average

of I2T and T2I MAP scores) system performance is best in the case of ZM image features, TFIDF text features, and Hebbian learning. This analysis aims to show the importance of choosing the methods wisely, as per the required operation and the application area. The performance (based upon the MAP score values) of various utilized methods can be visualized in the form of a bar chart in Figure 6.31.

6.3 Conclusion

After performing various experiments on the public Wikipedia data and primary endoscopy data, it has been deduced that the proposed cross-modal approaches $HSOM\_HEBB$ and $HSOM\_OJA$ outperform the state-of-the-art methods. Now, these approaches can be utilized for an effective cross-modal or multi-modal retrieval process using different modalities and in diverse application areas. Comparative analysis of $HSOM\_HEBB$ and $HSOM\_OJA$ approaches has also been shown using different image and text features on the endoscopy dataset.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

From the extensive review on cross-modal information retrieval presented in this thesis, it has been found that cross-modal systems are better than classic uni-modal systems in retrieving the multi-modal data and adding values to complement meaningful information. The dissertation 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. Miscellaneous applications in the field of cross-modal retrieval are mentioned. Challenges and open issues have also been discussed to help the research community in further research.

The proposed research in this study introduced new ways of intelligently training neural computing systems and querying them using images or text to retrieve matched texts or images respectively. The experimentation has been performed on the famous public Wikipedia dataset and the primary endoscopy dataset. In case of Wikipedia data, the visual features extracted from images are Zernike moments (ZM) that have almost no redundancy and LDA features are considered as the linguistic features for the text.

In case of endoscopy data, images are represented using ZM features and VGG16 deep features, and TFIDF and LDA features are extracted from the collateral text. Two unsupervised traditional self-organizing feature maps are trained simultaneously but separately for images and collateral text respectively. A Hebb link or Oja link is set up between the most active nodes in the two SOMs. This is the basis of our claim that we use multi-modal features for training neural networks and also establish cross-modal links between the two maps using an unsupervised Hebbian network while the training process. In reality, getting a labeled data is quite difficult, so the proposed framework will work effectively in that case as it is of unsupervised nature and thus does not require any data labeling. Experimentation, results, and the comparison with the state-of-the-art techniques prove the efficacy of the proposed technique in the field of cross-modal retrieval.

7.2 Research limitations

A few limitations of the proposed framework are given below:

1. Image and text vectorization techniques include few errors and approximations which could limit the performance of the system for the given data distribution. (i) For *image vectorization*, there are three types of errors involved in the calculation of Zernike moments [186]: (a) *Geometric error* due to mapping of a digital image into a unit circle with pixels; (b) *Discretization error* due to computer's digital representation of continuous variables; and (c) *Numerical integration* error due to the calculation of double integration through double summations while the centre of a grid is used to calculate the basis function; and (2) for *text vectorization*, vector size, uniqueness analysis, and topic selection for LDA remains an open question.

2. In the proposed study, only MAP score and accuracy measure have been considered to make the discussion simpler. Other measures such as precision and recall curve, F1-score, etc. should also be considered as per the problem requirement.

3. Default values are used for SOM training parameters such as neighborhood size and distance metric for node distance calculation.

7.3 Future scope

Prominent SOM parameters in this study have been selected after several experiments, however, a suitable technique should be used for SOM parameter tuning. Although the results obtained using the proposed approach are promising, image semantics must be considered more carefully for better performance. Diverse image and textual noise removal techniques should be considered in the future for further improvement in the MAP scores. The presented framework can be applied in miscellaneous application areas by utilizing suitable feature extractors based on the type of modality.

Modalities other than image and text can also be employed for experimentation using the proposed approach. Moreover, modified versions of SOM, Hebbian, and Oja learning rules can be incorporated into this study for further performance enhancement. The proposed method can be scaled by experimenting with large datasets. Different feature selection techniques can be applied to retrieve the best features representing modalities and obtain better results.

References

[1] G. Wang, H. Ji, D. Kong, and N. Zhang, "Modality-dependent cross-modal retrieval based on graph regularization," *Mobile Information Systems*, vol. 2020, pp. 1–17, 2020.

[2] H. T. Shen, L. Liu, Y. Yang, X. Xu, Z. Huang, F. Shen, and R. Hong, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[3] T. Nanda, B. Sahoo, and C. Chatterjee, "Enhancing the applicability of Kohonen self-organizing map (KSOM) estimator for gap-filling in hydrometeorological timeseries data," *Journal of Hydrology*, vol. 549, pp. 133–147, 2017.

[4] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[5] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[6] B. E. Stein, T. R. Stanford, and B. A. Rowland, "Development of multisensory integration from the perspective of the individual neuron," *Nature Reviews Neuroscience*, vol. 15, no. 8, pp. 520–535, 2014.

[7] R. L. Miller and B. A. Rowland, "Multisensory integration: How the brain combines information across the senses," *Computational Models of Brain and Behavior*, pp. 215–228, 2017.

[8] B. E. Stein and M. A. Meredith, *The Merging of the Senses.* The MIT Press, 1993.

[9] B. E. Stein, M. A. Meredith, W. S. Huneycutt, and L. McDade, "Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli," *Journal of Cognitive Neuroscience*, vol. 1, no. 1, pp. 12–24, 1989.

[10] A. Sen, S. Parida, K. Kotwal, S. Panda, O. Bojar, and S. R. Dash, "Bengali visual genome: A multimodal dataset for machine translation and image captioning," in *Intelligent Data Engineering and Analytics*, pp. 63–70, Springer, 2022.

[11] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end generative pretraining for multimodal video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17959–17968, 2022.

[12] L. V. B. Beltrán, J. C. Caicedo, N. Journet, M. Coustaty, F. Lecellier, and A. Doucet, "Deep multimodal learning for cross-modal retrieval: One model for all tasks," *Pattern Recognition Letters*, vol. 146, pp. 38–45, 2021.

[13] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 13059–13076, 2021.

[14] W. Zhang, J. Yu, W. Zhao, and C. Ran, "DMRFNet: deep multimodal reasoning and fusion for visual question answering and explanation generation," *Information Fusion*, vol. 72, pp. 70–79, 2021.

[15] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 6, pp. 1–18, 2021.

[16] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2014.

[17] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[18] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.

[19] C. Saraceno and R. Leonardi, "Indexing audiovisual databases through joint audio and video processing," *International Journal of Imaging Systems and Technology*, vol. 9, no. 5, pp. 320–331, 1998.

[20] 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, pp. 2369–2372, IEEE, 2000.

[21] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[22] T. Westerveld, D. Hiemstra, and F. De Jong, "Extracting bimodal representations for language-based image retrieval," in *Multimedia'99*, pp. 33–42, Springer, 2000.

[23] T. Westerveld, "Image retrieval: Content versus context.," in *Recherche d'Informations Assistée par Ordinateur*, pp. 276–284, Citeseer, 2000.

[24] K. Allen, "YouTube mistakenly flags Notre Dame Cathedral fire videos as 9/11 conspiracy." `https://abcnews.go.com/Business/youtube-mistakenly-flags-notre-dame-fire-videos-911/story?id=62419884`. Accessed: May 23, 2022.

[25] B. News, "Google apologises for Photos app's racist blunder." `https://www.bbc.com/news/technology-33347866`. Accessed: May 23, 2022.

[26] X. Xu, L. He, A. Shimada, R.-i. Taniguchi, and H. Lu, "Learning unified binary codes for cross-modal retrieval via latent semantic hashing," *Neurocomputing*, vol. 213, pp. 191–203, 2016.

[27] K. Ahmad, "Slandail: A security system for language and image analysis-project no: 607691," *Available at SSRN 3060047*, 2017.

[28] A. Hanbury, "A survey of methods for image annotation," *Journal of Visual Languages & Computing*, vol. 19, no. 5, pp. 617–627, 2008.

[29] B. Rafkind, M. Lee, S.-F. Chang, and 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*, pp. 73–80, Association for Computational Linguistics, 2006.

[30] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1367–1374, IEEE, 2009.

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

[32] J. Marin, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba, "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 187–203, 2019.

[33] H. H. Lee, K. Shu, P. Achananuparp, P. K. Prasetyo, Y. Liu, E.-P. Lim, and L. R. Varshney, "Recipegpt: Generative pre-training based cooking recipe generation and evaluation system," in *Companion Proceedings of the Web Conference 2020*, pp. 181–184, 2020.

[34] V. H. Contreras, J. S. Lara, O. J. Perdomo, and F. A. González, "Supervised online matrix factorization for histopathological multimodal retrieval," in *14th International Symposium on Medical Information Processing and Analysis*, vol. 10975, p. 109750Y, International Society for Optics and Photonics, 2018.

[35] N. Garcia and G. Vogiatzis, "How to read paintings: semantic art understanding with multi-modal retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

[36] N. Garcia, B. Renoust, and Y. Nakashima, "Context-aware embeddings for automatic art analysis," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 25–33, 2019.

[37] M. Jing, B. W. Scotney, S. A. Coleman, M. T. McGinnity, X. Zhang, S. Kelly, K. Ahmad, A. Schlaf, S. Gründer-Fahrer, and G. Heyer, "Integration of text and image analysis for flood event image recognition," in *2016 27th Irish Signals and Systems Conference (ISSC)*, pp. 1–6, IEEE, 2016.

[38] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2018.

[39] M. M. Rahman, D. You, M. S. Simpson, S. K. Antani, D. Demner-Fushman, and G. R. Thoma, "Interactive cross and multimodal biomedical image retrieval based on automatic region-of-interest (ROI) identification and classification," *International Journal of Multimedia Information Retrieval*, vol. 3, no. 3, pp. 131–146, 2014.

[40] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*, pp. 162–190, Springer, 1992.

[41] C. Guo and D. Wu, "Canonical correlation analysis (CCA) based multi-view learning: An overview," *arXiv preprint arXiv:1907.01693*, 2019.

[42] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[43] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 251–260, 2010.

[44] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 521–535, 2013.

[45] Y. Verma and C. Jawahar, "Im2Text and Text2Im: Associating images and texts for cross-modal retrieval.," in *British Machine Vision Conference*, vol. 1, p. 2, Citeseer, 2014.

[46] Y. Verma and C. Jawahar, "A support vector approach for cross-modal search of images and texts," *Computer Vision and Image Understanding*, vol. 154, pp. 48–63, 2017.

[47] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 1059–1074, 2014.

[48] J. Shao, Z. Zhao, F. Su, and T. Yue, "Towards improving canonical correlation analysis for cross-modal retrieval," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 332–339, 2017.

[49] W. Xiong, S. Wang, C. Zhang, and 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)*, pp. 1–6, IEEE, 2013.

[50] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4094–4102, 2015.

[51] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval.," in *British Machine Vision Conference*, vol. 1, p. 5, 2010.

[52] S. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *International Journal of Computer Vision*, vol. 100, no. 2, pp. 134–153, 2012.

[53] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2010–2023, 2015.

[54] G. Xu, X. Li, and Z. Zhang, "Semantic consistency cross-modal retrieval with semi-supervised graph regularization," *IEEE Access*, vol. 8, pp. 14278–14288, 2020.

[55] G. Xu, X. Li, L. Shi, Z. Zhang, and A. Zhai, "Combination subspace graph learning for cross-modal retrieval," *Alexandria Engineering Journal*, 2020.

[56] M. Zhang, H. Zhang, J. Li, L. Wang, Y. Fang, and J. Sun, "Supervised graph regularization based cross media retrieval with intra and inter-class correlation," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 1–11, 2019.

[57] M. Zhang, H. Zhang, J. Li, Y. Fang, L. Wang, and F. Shang, "Multi-modal graph regularization based class center discriminant analysis for cross modal retrieval," *Multimedia Tools and Applications*, vol. 78, no. 19, pp. 28285–28307, 2019.

[58] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 128–141, 2017.

[59] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, and S. Yan, "Modality-dependent cross-media retrieval," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 4, pp. 1–13, 2016.

[60] C. Deng, X. Tang, J. Yan, W. Liu, and X. Gao, "Discriminative dictionary learning with common label alignment for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 208–218, 2015.

[61] S. Wang, F. Zhuang, S. Jiang, Q. Huang, and Q. Tian, "Cluster-sensitive structured correlation analysis for web cross-modal retrieval," *Neurocomputing*, vol. 168, pp. 747–760, 2015.

[62] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Cross-modal retrieval using multiordered discriminative structured subspace learning," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1220–1233, 2016.

[63] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 154–162, 2017.

[64] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2542–2555, 2017.

[65] Y. Wu, S. Wang, G. Song, and Q. Huang, "Augmented adversarial training for cross-modal retrieval," *IEEE Transactions on Multimedia*, 2020.

[66] F. Shang, H. Zhang, L. Zhu, and J. Sun, "Adversarial cross-modal retrieval based on dictionary learning," *Neurocomputing*, vol. 355, pp. 93–104, 2019.

[67] J. Jeon, V. Lavrenko, and 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 Information Retrieval*, pp. 119–126, 2003.

[68] Y. Xia, Y. Wu, and J. Feng, "Cross-media retrieval using probabilistic model of automatic image annotation," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 4, pp. 145–154, 2015.

[69] Z. Li, J. Liu, C. Xu, and H. Lu, "Mlrank: Multi-correlation learning to rank for image annotation," *Pattern Recognition*, vol. 46, no. 10, pp. 2700–2710, 2013.

[70] Q. Xu, M. Li, and M. Yu, "Learning to rank with relational graph and point-wise constraint for cross-modal retrieval," *Soft Computing*, vol. 23, no. 19, pp. 9413–9427, 2019.

[71] Y. Wu, S. Wang, and Q. Huang, "Online fast adaptive low-rank similarity learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, 2019.

[72] J. Yu, Y. Cong, Z. Qin, and T. Wan, "Cross-modal topic correlations for multimedia retrieval," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 246–249, IEEE, 2012.

[73] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang, "Multi-modal mutual topic reinforce modeling for cross-media retrieval," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 307–316, 2014.

[74] Z. Qin, J. Yu, Y. Cong, and T. Wan, "Topic correlation model for cross-modal multimedia information retrieval," *Pattern Analysis and Applications*, vol. 19, no. 4, pp. 1007–1022, 2016.

[75] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[76] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *International Conference on Machine Learning Workshop*, vol. 79, 2012.

[77] D. Xia, L. Miao, and A. Fan, "A cross-modal multimedia retrieval method using depth correlation mining in big data environment," *Multimedia Tools and Applications*, pp. 1–16, 2019.

[78] B. Jiang, J. Yang, Z. Lv, K. Tian, Q. Meng, and Y. Yan, "Internet cross-media retrieval based on deep learning," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 356–366, 2017.

[79] P. Hu, L. Zhen, D. Peng, and 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*, pp. 635–644, 2019.

[80] F. Feng, X. Wang, R. Li, and I. Ahmad, "Correspondence autoencoders for cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 1s, p. 26, 2015.

[81] D. Mandal, P. Rao, and S. Biswas, "Semi-supervised cross-modal retrieval with label prediction," *IEEE Transactions on Multimedia*, 2019.

[82] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning*, pp. 595–603, 2014.

[83] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 7–16, 2014.

[84] F. Feng, R. Li, and X. Wang, "Deep correspondence restricted boltzmann machine for cross-modal retrieval," *Neurocomputing*, vol. 154, pp. 50–60, 2015.

[85] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 449–460, 2016.

[86] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

[87] X. Huang, Y. Peng, and M. Yuan, "Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Transactions on Cybernetics*, 2018.

[88] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and 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*, pp. 35–44, 2018.

[89] J. Gu, J. Cai, S. R. Joty, L. Niu, and 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*, pp. 7181–7189, 2018.

[90] W. Cao, Q. Lin, Z. He, and Z. He, "Hybrid representation learning for cross-modal retrieval," *Neurocomputing*, vol. 345, pp. 45–57, 2019.

[91] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, 2019.

[92] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2400–2413, 2019.

[93] Z. Yang, Z. Lin, P. Kang, J. Lv, Q. Li, and W. Liu, "Learning shared semantic space with correlation alignment for cross-modal event retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1, pp. 1–22, 2020.

[94] B. Zhang, L. Zhu, J. Sun, and H. Zhang, "Cross-media retrieval with collective deep semantic learning," *Multimedia Tools and Applications*, vol. 77, no. 17, pp. 22247–22266, 2018.

[95] R. Shriwas, P. Joshi, V. M. Ladwani, and V. Ramasubramanian, "Multimodal associative storage and retrieval using Hopfield auto-associative memory network," in *International Conference on Artificial Neural Networks*, pp. 57–75, Springer, 2019.

[96] J.-W. Ha, B.-H. Kim, B. Lee, and B.-T. Zhang, "Layered hypernetwork models for cross-modal associative text and image keyword generation in multimodal information retrieval," in *Pacific Rim International Conference on Artificial Intelligence*, pp. 76–87, Springer, 2010.

[97] J.-W. Ha, B.-J. Lee, and B.-T. Zhang, "Text-to-image retrieval based on incremental association via multimodal hypernetworks," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3245–3250, IEEE, 2012.

[98] Z. Liu and X. Wang, "Cross-modal associative memory by MultiSOM," in *2014 4th International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE)*, pp. 1–5, IEEE, 2014.

[99] S. Wermter, C. Weber, and M. Elshaw, "Associative neural models for biomimetic multi-modal learning in a mirror neuron-based robot," in *Modeling Language, Cognition and Action*, pp. 31–46, World Scientific, 2005.

[100] G. Collell, T. Zhang, and M.-F. Moens, "Learning to predict: A fast reconstructive method to generate multimodal embeddings," *arXiv preprint arXiv:1703.08737*, 2017.

[101] S. Wang, J. Zhang, and C. Zong, "Associative multichannel autoencoder for multimodal word representation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 115–124, 2018.

[102] J.-H. Su, C.-L. Chou, C.-Y. Lin, and V. S. Tseng, "Effective semantic annotation by image-to-concept distribution model," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 530–538, 2011.

[103] J. C. Pereira and N. Vasconcelos, "Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems," *Computer Vision and Image Understanding*, vol. 124, pp. 123–135, 2014.

[104] L. Wang, L. Zhu, E. Yu, J. Sun, and H. Zhang, "Task-dependent and query-dependent subspace learning for cross-modal retrieval," *IEEE Access*, vol. 6, pp. 27091–27102, 2018.

[105] M. Xu, Z. Zhu, Y. Zhao, and F. Sun, "Subspace learning by kernel dependence maximization for cross-modal retrieval," *Neurocomputing*, vol. 309, pp. 94–105, 2018.

[106] E. Yu, J. Li, L. Wang, J. Zhang, W. Wan, and J. Sun, "Multi-class joint subspace learning for cross-modal retrieval," *Pattern Recognition Letters*, vol. 130, pp. 165–173, 2020.

[107] X. Dong, J. Sun, P. Duan, L. Meng, Y. Tan, W. Wan, H. Wu, B. Zhang, and H. Zhang, "Semi-supervised modality-dependent cross-media retrieval," *Multimedia Tools and Applications*, vol. 77, no. 3, pp. 3579–3595, 2018.

[108] L. Chi and X. Zhu, "Hashing techniques: A survey and taxonomy," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–36, 2017.

[109] H. P. Luhn, "A new method of recording and searching information," *American Documentation*, vol. 4, no. 1, pp. 14–16, 1953.

[110] H. Stevens, "Hans Peter Luhn and the birth of the hashing algorithm," *IEEE Spectrum*, vol. 55, no. 2, pp. 44–49, 2018.

[111] W. W. Peterson, "Addressing for random-access storage," *IBM Journal of Research and Development*, vol. 1, no. 2, pp. 130–146, 1957.

[112] R. Morris, "Scatter storage techniques," *Communications of the ACM*, vol. 11, no. 1, pp. 38–44, 1968.

[113] L. Xie, L. Zhu, P. Pan, and Y. Lu, "Cross-modal self-taught hashing for large-scale image retrieval," *Signal Processing*, vol. 124, pp. 81–92, 2016.

[114] W. Cao, W. Feng, Q. Lin, G. Cao, and Z. He, "A review of hashing methods for multimodal retrieval," *IEEE Access*, vol. 8, pp. 15377–15391, 2020.

[115] Y. Fang, H. Zhang, and Y. Ren, "Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing," *Knowledge-Based Systems*, vol. 171, pp. 69–80, 2019.

[116] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 143–152, 2013.

[117] B. Wu, Q. Yang, W.-S. Zheng, Y. Wang, and J. Wang, "Quantized correlation hashing for fast cross-modal search," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[118] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018.

[119] C. Yan, X. Bai, S. Wang, J. Zhou, and E. R. Hancock, "Cross-modal hashing with semantic deep embedding," *Neurocomputing*, vol. 337, pp. 58–66, 2019.

[120] X. Lu, L. Zhu, Z. Cheng, X. Song, and H. Zhang, "Efficient discrete latent semantic hashing for scalable cross-modal retrieval," *Signal Processing*, vol. 154, pp. 217–231, 2019.

[121] Y. Cao, M. Long, J. Wang, Q. Yang, and 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*, pp. 1445–1454, 2016.

[122] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3232–3240, 2017.

[123] J. Yu, X.-J. Wu, and J. Kittler, "Learning discriminative hashing codes for cross-modal retrieval based on multi-view features," *Pattern Analysis and Applications*, pp. 1–18, 2020.

[124] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3157–3166, 2016.

[125] X. Liu, Z. Li, J. Wang, G. Yu, C. Domeniconi, and X. Zhang, "Cross-modal zero-shot hashing," *arXiv preprint arXiv:1908.07388*, 2019.

[126] J. Yu and X.-J. Wu, "Unsupervised concatenation hashing with sparse constraint for cross-modal retrieval," *arXiv preprint arXiv:1904.00726*, 2019.

[127] X. Zhang, H. Lai, and J. Feng, "Attention-aware deep adversarial hashing for cross-modal retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 591–606, 2018.

[128] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2916–2929, 2012.

[129] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[130] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in Neural Information Processing Systems*, pp. 1753–1760, 2009.

[131] J. Song, Y. Yang, Y. Yang, Z. Huang, and 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*, pp. 785–796, 2013.

[132] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7380–7388, 2017.

[133] X. Shen, F. Shen, Q.-S. Sun, Y.-H. Yuan, and H. T. Shen, "Robust cross-view hashing for multimedia retrieval," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 893–897, 2016.

[134] J. Zhou, G. Ding, and 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*, pp. 415–424, 2014.

[135] Z. Ji, W. Yao, W. Wei, H. Song, and H. Pi, "Deep multi-level semantic hashing for cross-modal retrieval," *IEEE Access*, vol. 7, pp. 23667–23674, 2019.

[136] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proceedings of ACM Conference on Image and Video Retrieval (CIVR'09)*, (Santorini, Greece.), July 8-10, 2009.

[137] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems," in *International Workshop OntoImage*, vol. 2, 2006.

[138] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[139] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, 2008.

[140] M. J. Huiskes, B. Thomee, and 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*, pp. 527–536, 2010.

[141] J. Krapac, M. Allan, J. Verbeek, and F. Juried, "Improving web image search results using query-relative classifiers," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1094–1101, IEEE, 2010.

[142] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[143] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[144] C. Rashtchian, P. Young, M. Hodosh, and 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*, pp. 139–147, Association for Computational Linguistics, 2010.

[145] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, pp. 740–755, Springer, 2014.

[146] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Ieee, 2009.

[147] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *2011 International Conference on Computer Vision*, pp. 2407–2414, IEEE, 2011.

[148] F. Zhong, G. Wang, Z. Chen, F. Xia, and G. Min, "Cross-modal retrieval for CPSS data," *IEEE Access*, vol. 8, pp. 16689–16701, 2020.

[149] Y. Wang, X. Lin, L. Wu, W. Zhang, and 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*, pp. 999–1002, 2015.

[150] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5427–5440, 2016.

[151] X. Zhang and K. Ahmad, "Ontology and terminology of disaster management," in *DIMPLE: DIsaster Management and Principled Large-scale information Extraction Workshop Programme*, p. 46, 2014.

[152] K. Ahmad and M. Rogers, "Corpus linguistics and terminology extraction," in *Handbook of Terminology Management (Volume 2)* (S. E. Wright and G. Budin, eds.), pp. 725–760, Amsterdam and Philadelphia: John Benjamins Publishing Company, 2001.

[153] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 370–381, 2015.

[154] N. Gao, S.-J. Huang, Y. Yan, and S. Chen, "Cross modal similarity learning with active queries," *Pattern Recognition*, vol. 75, pp. 214–222, 2018.

[155] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1602–1612, 2018.

[156] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks.," in *International Joint Conference on Artificial Intelligence*, pp. 3846–3853, 2016.

[157] J. Shao, L. Wang, Z. Zhao, A. Cai, *et al.*, "Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval," *Neurocomputing*, vol. 214, pp. 618–628, 2016.

[158] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Deep coupled metric learning for cross-modal matching," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1234–1244, 2016.

[159] J. Luo, Y. Shen, X. Ao, Z. Zhao, and M. Yang, "Cross-modal image-text retrieval with multitask learning," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2309–2312, 2019.

[160] Y. Jian, J. Xiao, Y. Cao, A. Khan, and J. Zhu, "Deep pairwise ranking with multi-label information for cross-modal retrieval," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1810–1815, IEEE, 2019.

[161] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 839–847, 2017.

[162] C. Tian, V. De Silva, M. Caine, and S. Swanson, "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data," *Applied Sciences*, vol. 10, no. 1, p. 24, 2020.

[163] D. J. Armaghani, G. D. Hatzigeorgiou, C. Karamani, A. Skentou, I. Zoumpoulaki, and P. G. Asteris, "Soft computing-based techniques for concrete beams shear strength," *Procedia Structural Integrity*, vol. 17, pp. 924–933, 2019.

[164] C. Raghuraman, S. Suresh, S. Shivshankar, and R. Chapaneri, "Static and dynamic malware analysis using machine learning," in *First International Conference on Sustainable Technologies for Computational Intelligence*, pp. 793–806, Springer, 2020.

[165] H. Müller and D. Unay, "Retrieval from and understanding of large-scale multi-modal medical datasets: A review," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2093–2104, 2017.

[166] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[167] A. C. Duarte, "Cross-modal neural sign language translation," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1650–1654, ACM, 2019.

[168] Y. Jia, L. Bai, S. Liu, P. Wang, J. Guo, and Y. Xie, "Semantically-enhanced kernel canonical correlation analysis: a multi-label cross-modal retrieval," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13169–13188, 2019.

[169] A. J. Fredo, R. Abilash, R. Femi, A. Mythili, and C. S. Kumar, "Classification of damages in composite images using Zernike moments and support vector machines," *Composites Part B: Engineering*, vol. 168, pp. 77–86, 2019.

[170] B. Kaur, S. Singh, and J. Kumar, "Iris recognition using Zernike moments and polar harmonic transforms," *Arabian Journal for Science and Engineering*, vol. 43, no. 12, pp. 7209–7218, 2018.

[171] H. Aggarwal and D. K. Vishwakarma, "Covariate conscious approach for Gait recognition based upon Zernike moment invariants," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 397–407, 2017.

[172] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[173] L. Zheng, Z. Caiming, and C. Caixian, "MMDF-LDA: An improved multi-modal Latent Dirichlet Allocation model for social image annotation," *Expert Systems with Applications*, vol. 104, pp. 168–184, 2018.

[174] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter," *PloS one*, vol. 15, no. 9, p. e0239441, 2020.

[175] P. Kaur, H. S. Pannu, and A. K. Malhi, "Plant disease recognition using fractional-order Zernike moments and SVM classifier," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8749–8768, 2019.

[176] H. Shu, L. Luo, and J.-l. Caatrieux, "Moment-based approaches in imaging. 1. basic features [a look at...]," *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, no. 5, pp. 70–74, 2007.

[177] J. Flusser, B. Zitova, and T. Suk, *Moments and Moment Invariants in Pattern Recognition*. John Wiley & Sons, 2009.

[178] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comprehensive study of continuous orthogonal moments-a systematic review," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–30, 2019.

[179] S. Azzouzi, A. Hjouji, J. EL-Mekkaoui, and A. EL Khalfi, "An improved image clustering algorithm based on Kernel method and Tchebychev orthogonal moments," *Evolutionary Intelligence*, pp. 1–22, 2022.

[180] R. D. C. da Silva, T. R. Jenkyn, and V. A. Carranza, "Enhanced preprocessing for deep learning in MRI whole brain segmentation using orthogonal moments," *Brain Multiphysics*, vol. 3, p. 100049, 2022.

[181] S. H. Abdulhussain, B. M. Mahmmod, M. A. Naser, M. Q. Alsabah, R. Ali, and S. Al-Haddad, "A robust handwritten numeral recognition using hybrid orthogonal polynomials and moments," *Sensors*, vol. 21, no. 6, p. 1999, 2021.

[182] F. Akhmedova and S. Liao, "Face recognition with discrete orthogonal moments," in *Recent Advances in Computer Vision*, pp. 189–209, Springer, 2019.

[183] Z. von F, "Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode," *Physica*, vol. 1, no. 7-12, pp. 689–704, 1934.

[184] A. Aggarwal and C. Singh, "Zernike moments-based Gurumukhi character recognition," *Applied Artificial Intelligence*, vol. 30, no. 5, pp. 429–444, 2016.

[185] M. R. Teague, "Image analysis via the general theory of moments," *Josa*, vol. 70, no. 8, pp. 920–930, 1980.

[186] A. Khotanzad and Y. H. Hong, "Invariant image recognition by Zernike moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 489–497, 1990.

[187] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[188] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[189] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019.

[190] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[191] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, 2013.

[192] M. Pacella, A. Grieco, and M. Blaco, "On the use of self-organizing map for text clustering in engineering change process analysis: a case study," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.

[193] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press, 2005.

[194] J. Zhou, X. Fu, S. Zhou, J. Zhou, H. Ye, and H. T. Nguyen, "Automated segmentation of soybean plants from 3D point cloud using machine learning," *Computers and Electronics in Agriculture*, vol. 162, pp. 143–153, 2019.

[195] J. Zhang, Q. Wu, J. Zhang, C. Shen, J. Lu, and Q. Wu, "Heritage image annotation via collective knowledge," *Pattern Recognition*, vol. 93, pp. 204–214, 2019.

[196] E. Aghajari and G. D. Chandrashekhar, "Self-organizing map based extended fuzzy C-means (SEEFC) algorithm for image segmentation," *Applied Soft Computing*, vol. 54, pp. 347–363, 2017.

[197] V. R. Borges, M. C. F. de Oliveira, T. G. Silva, A. A. H. Vieira, and B. Hamann, "Region growing for segmenting green microalgae images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 1, pp. 257–270, 2016.

[198] W. Xue-guang and C. Shu-hong, "An improved image segmentation algorithm based on two-dimensional Otsu method," *Information Sciences Letters*, vol. 1, no. 2, pp. 77–83, 2012.

[199] K. Muhammad, S. Khan, N. Kumar, J. Del Ser, and S. Mirjalili, "Vision-based personalized wireless capsule endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges," *Future Generation Computer Systems*, vol. 113, pp. 266–280, 2020.

[200] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 831–839, 2020.

[201] Y. Cong, S. Wang, J. Liu, J. Cao, Y. Yang, and J. Luo, "Deep sparse feature selection for computer aided endoscopy diagnosis," *Pattern Recognition*, vol. 48, no. 3, pp. 907–917, 2015.

[202] K. Radhika, M. Venkatesha, and G. Sekhar, "An approach for on-line signature authentication using Zernike moments," *Pattern Recognition Letters*, vol. 32, no. 5, pp. 749–760, 2011.

[203] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5147–5156, 2016.

[204] W. Zhu, J. Lu, and J. Zhou, "Nonlinear subspace clustering for image clustering," *Pattern Recognition Letters*, vol. 107, pp. 131–136, 2018.

[205] T. Yellamraju and M. Boutin, "Clusterability and clustering of images and other "real" high-dimensional data," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1927–1938, 2018.

[206] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5879–5887, 2017.

[207] S. Park, S. Han, S. Kim, D. Kim, S. Park, S. Hong, and M. Cha, "Improving unsupervised image clustering with robust learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12278–12287, 2021.

[208] Y. Ren, N. Wang, M. Li, and Z. Xu, "Deep density-based image clustering," *Knowledge-Based Systems*, vol. 197, p. 105841, 2020.

[209] W. Hu, C. Chen, F. Ye, Z. Zheng, and Y. Du, "Learning deep discriminative representations with pseudo supervision for image clustering," *Information Sciences*, vol. 568, pp. 199–215, 2021.

[210] A. Kar, S. Pramanik, A. Chakraborty, D. Bhattacharjee, E. S. Ho, and H. P. Shum, "LMZMPM: local modified Zernike moment per-unit mass for robust human face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 495–509, 2020.

[211] M. Khare and A. Khare, "Integration of complex wavelet transform and Zernike moment for multi-class classification," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 1151–1162, 2021.

[212] C. Singh and A. Aggarwal, "An efficient approach for image sequence denoising using Zernike moments-based nonlocal means approach," *Computers & Electrical Engineering*, vol. 62, pp. 330–344, 2017.

[213] X. Fan and T. Tjahjadi, "A dynamic framework based on local Zernike moment and motion history image for facial expression recognition," *Pattern recognition*, vol. 64, pp. 399–406, 2017.

[214] J. Yang and J.-y. Yang, "From image vector to matrix: A straightforward image projection technique—IMPCA vs. PCA," *Pattern Recognition*, vol. 35, no. 9, pp. 1997–1999, 2002.

[215] A. Patel, K. Rani, S. Kumar, I. N. Figueiredo, and P. N. Figueiredo, "Automated bleeding detection in wireless capsule endoscopy images based on sparse coding," *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30353–30366, 2021.

[216] V. Gupta, T. Priya, A. K. Yadav, R. B. Pachori, and U. R. Acharya, "Automated detection of focal EEG signals using features extracted from flexible analytic wavelet transform," *Pattern Recognition Letters*, vol. 94, pp. 180–188, 2017.

[217] C. Kurtz, A. Depeursinge, S. Napel, C. F. Beaulieu, and D. L. Rubin, "On combining image-based and ontological semantic dissimilarities for medical image retrieval applications," *Medical Image Analysis*, vol. 18, no. 7, pp. 1082–1100, 2014.

[218] E. Alegre, V. González-Castro, R. Alaiz-Rodríguez, and M. T. García-Ordás, "Texture and moments-based classification of the acrosome integrity of boar spermatozoa images," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 2, pp. 873–881, 2012.

[219] A. K. Kundu and S. A. Fattah, "Probability density function based modeling of spatial feature variation in capsule endoscopy data for automatic bleeding detection," *Computers in Biology and Medicine*, vol. 115, p. 103478, 2019.

[220] Y. Yuan and M. Q.-H. Meng, "Deep learning for polyp recognition in wireless capsule endoscopy images," *Medical Physics*, vol. 44, no. 4, pp. 1379–1389, 2017.

[221] D. Zhang, "Wavelet transform," in *Fundamentals of Image Data Mining*, pp. 35–44, Springer, 2019.

[222] A.-W. Deng, C.-H. Wei, and C.-Y. Gwo, "Stable, fast computation of high-order Zernike moments using a recursive method," *Pattern Recognition*, vol. 56, pp. 16–25, 2016.

[223] S. A. A. Karim, M. H. Kamarudin, B. A. Karim, M. K. Hasan, and J. Sulaiman, "Wavelet transform and fast Fourier transform for signal compression: A comparative study," in *2011 International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, pp. 280–285, IEEE, 2011.

[224] I. Daubechies, *Different Perspectives on Wavelets*, vol. 47. American Mathematical Soc., 2016.

[225] C.-H. Hsia, J.-S. Chiang, J.-M. Guo, and H. Olkkonen, "Multiple moving objects detection and tracking using discrete wavelet transform," in *Discrete Wavelet Transforms-Biomedical Applications*, pp. 297–320, IntechOpen, 2011.

[226] A. Krishnaswamy Rangarajan and R. Purushothaman, "Disease classification in eggplant using pre-trained VGG16 and MSVM," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[227] H. Yang, J. Ni, J. Gao, Z. Han, and T. Luan, "A novel method for peanut variety identification and classification by improved VGG16," *Scientific Reports*, vol. 11, no. 1, pp. 1–17, 2021.

[228] E. Rezende, G. Ruppert, T. Carvalho, A. Theophilo, F. Ramos, and P. d. Geus, "Malicious software classification using VGG16 deep neural network's bottleneck features," in *Information Technology-New Generations*, pp. 51–59, Springer, 2018.

[229] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[230] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.

[231] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of Mathematical Biology*, vol. 15, no. 3, pp. 267–273, 1982.

[232] L. Xie, L. Zhu, and G. Chen, "Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval," *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 9185–9204, 2016.

[233] H. Liu, F. Wang, X. Zhang, and F. Sun, "Weakly-paired deep dictionary learning for cross-modal retrieval," *Pattern Recognition Letters*, vol. 130, pp. 199–206, 2020.

[234] G. Doyle and C. Elkan, "Accounting for burstiness in topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 281–288, 2009.

[235] S. Lokesh, P. M. Kumar, M. R. Devi, P. Parthasarathy, and C. Gokulnath, "An automatic Tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map," *Neural Computing and Applications*, vol. 31, no. 5, pp. 1521–1531, 2019.

[236] J. Tervonen, S. Puttonen, M. J. Sillanpää, L. Hopsu, Z. Homorodi, J. Keränen, J. Pajukanta, A. Tolonen, A. Lämsä, and J. Mäntyjärvi, "Personalized mental stress detection with self-organizing map: From laboratory to the field," *Computers in Biology and Medicine*, vol. 124, p. 103935, 2020.

[237] M. Nilashi, H. Ahmadi, A. A. Manaf, T. A. Rashid, S. Samad, L. Shahmoradi, N. Aljojo, and E. Akbari, "Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates," *International Journal of Fuzzy Systems*, vol. 22, no. 4, pp. 1376–1388, 2020.

[238] Y. Chen, N. Ashizawa, C. K. Yeo, N. Yanai, and S. Yean, "Multi-scale self-organizing map assisted deep autoencoding gaussian mixture model for unsupervised intrusion detection," *Knowledge-Based Systems*, vol. 224, p. 107086, 2021.

[239] G. Amato, F. Carrara, F. Falchi, C. Gennaro, and G. Lagani, "Hebbian learning meets deep convolutional neural networks," in *International Conference on Image Analysis and Processing*, pp. 324–334, Springer, 2019.

[240] S. Bekhouche, F. Dornaika, A. Benlamoudi, A. Ouafi, and A. Taleb-Ahmed, "A comparative study of human facial age estimation: handcrafted features vs. deep features," *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 26605–26622, 2020.

[241] S. A. Siddiqui, A. Salman, M. I. Malik, F. Shafait, A. Mian, M. R. Shortis, and E. S. Harvey, "Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data," *ICES Journal of Marine Science*, vol. 75, no. 1, pp. 374–389, 2018.

[242] O. Ghorbanzadeh, T. Blaschke, K. Gholamnia, S. R. Meena, D. Tiede, and J. Aryal, "Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection," *Remote Sensing*, vol. 11, no. 2, p. 196, 2019.

[243] H. S. Pannu, S. Ahuja, N. Dang, S. Soni, and A. K. Malhi, "Deep learning based image classification for intestinal hemorrhage," *Multimedia Tools and Applications*, vol. 79, pp. 21941–21966, 2020.

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

[245] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.