Hybrid SOM based cross-modal retrieval exploiting Hebbian learning

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

[a] *Computer Science and Engineering Department,*
Thapar Institute of Engineering and Technology, Patiala, India
[b] *Department of Computer Science, Aalto University, Finland*
[c] *Department of Computing and Informatics, Bournemouth University, UK*

Abstract

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 retrieved from the other in an efficient way. So, a novel unsupervised cross-modal retrieval framework based on associative learning is proposed in this paper where two traditional SOMs are trained separately for images and collateral text and then they are associated together using the Hebbian learning network to facilitate the cross-modal retrieval process. Experimental outcomes on a popular Wikipedia dataset demonstrate that the presented technique outshines various existing state-of-the-art techniques.

Keywords: Self organizing maps, cross-modal retrieval, Hebbian learning, Zernike moments, machine learning

1. Introduction

In reality, data is usually represented in diverse forms or is composed of different domains. Hence, the data associated with the same underlying event, content, or object may exist in the form of different modalities and exhibit heterogeneous characteristics. For instance, while visiting a new place, we record the visit by recording a video, taking pictures, or posting a piece of micro-blog. All these data forms are different, however, present the same content. So, classic uni-modal (incorporating single modality data) information retrieval approaches are of the least use nowadays when a huge amount of multi-modal data is being produced every day. There is an immediate requirement of effective multi-modal or cross-modal (incorporating information from numerous modalities) data analysis and retrieval techniques. Figure (1) shows few examples depicting the cross-modal retrieval process where one form of data can be retrieved using another form of data e.g. images and/or videos from the text.

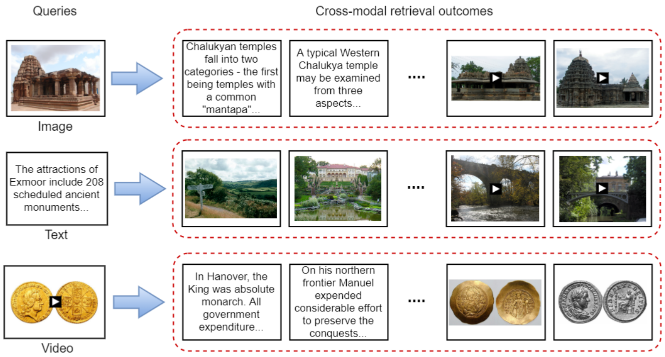

Queries
Cross-modal retrieval outcomes
Image
Chalukyan temples fall into two categories - the first being temples with a common "mantapa"...
A typical Western Chalukya temple may be examined from three aspects...
....
Text
The attractions of Exmoor include 208 scheduled ancient monuments...
....
Video
In Hanover, the King was absolute monarch. All government expenditure...
On his northern frontier Manuel expended considerable effort to preserve the conquests...
....

Figure 1: Few examples demonstrating the process of cross-modal retrieval

Humans familiarize themselves with the surroundings through various sensory modes where each mode provides a distinguishing impression of the environment [1]. Each sensory modality works individually to interconnect with the surroundings and obtain information, however, the knowledge acquired from all the modalities is fused inside the brain into a considerable awareness concerning the environment [2] (As shown in figure 2). For instance, if a person is unable to completely understand the meaning of what another person is saying then he will consequently start noticing body or facial expressions. Similarly, a single source of information is inadequate for a complete understanding of an incident or an object. Thus, a concept of information fusion comprising image and text is studied in this work 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 includes three unsupervised neural networks: (1) one SOM is trained to cluster images; (2) another SOM learns the text; and (3) the third unsupervised network (Hebbian network) links the highly active nodes on image SOM with nodes on text SOM (figure 3). The final system after merging both image and text SOM is known as hybrid SOM (HSOM) or multi-net system.

[*] Corresponding author

Email addresses: pkaur60_phd18@thapar.edu (Parminder Kaur), amalhi@bournemouth.ac.uk (Avleen Kaur Malhi), hspannu@thapar.edu (Husanbir Singh Pannu)

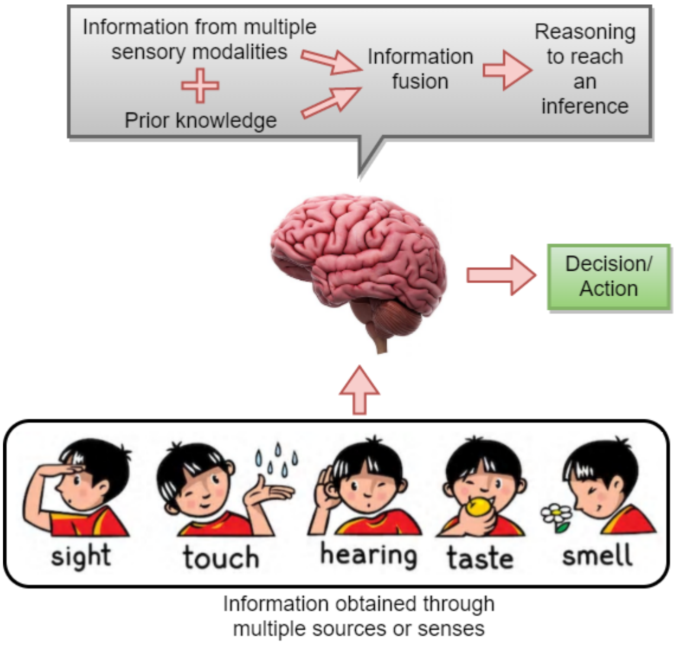

Information from multiple sensory modalities
Prior knowledge
Information fusion
Reasoning to reach an inference
Decision/ Action
sight
touch
hearing
taste
smell
Information obtained through multiple sources or senses

Figure 2: Simple illustration of information fusion inside brain

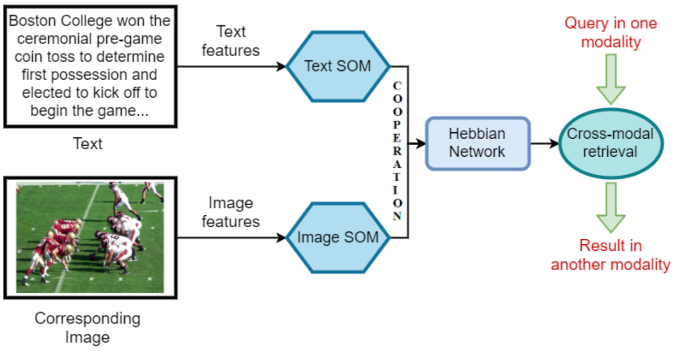

Boston College won the ceremonial pre-game coin toss to determine first possession and elected to kick off to begin the game...
Text
Text features
Text SOM
Image features
Image SOM
Corresponding Image
COOPERATION
Hebbian Network
Query in one modality
Cross-modal retrieval
Result in another modality

Figure 3: Flow diagram of the proposed cross-modal information retrieval system

Contributions

The prominent contributions of the article are given below:

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

2. Associative Hebbian learning technique has been utilized to link two SOMs which are separately trained using image and collateral text data.

3. The presented hybrid SOM framework is influenced by the working of the brain as diverse data representations are combined easily using an unsupervised Hebbian network.

The remaining article is organized as follows: *Section 2* presents the existing related techniques for image-text cross-modal retrieval, *Section 3* describes the extracted image and text features, *Section 4* is devoted to the introduced cross-modal method, experimental analysis is shown in *Section 5* which comprises of dataset, evaluation metrics, brief explanation of comparison methods, parameter settings, model training, and the results obtained, and *Section 6* provides the conclusion of the work.

2. Related Work

Numerous approaches have been introduced by the researchers to deal with the cross-modal issue in which one data form is retrieved from another data form. This section summarizes state-of-the-art works carried out in the area.

2.1. Cross-modal retrieval techniques

The major issue in multi-modal retrieval is the heterogeneity gap among different modalities. For dealing with this issue, most of the researchers have followed the approach of representing heterogeneous data in a common shared subspace where the modality gap can be reduced as much as possible. Various works proposed recently which deals with the cross-modal problem are summarized in this section.

2.1.1. Subspace learning based methods

Subspace learning methods have a crucial role in the area of cross-modal retrieval. Canonical correlation analysis (CCA) is one of the initial and popular subspace learning based approach introduced by *Hotelling* [3]. It is an unsupervised technique whose principal logic is to detect a pair of projections for different modalities so that the correlation between them is maximized [4]. In [5], authors have proposed CCA for correlation analysis, semantic matching (SM) is done for representing the image and textual data at a higher level of abstraction, and semantic correlation matching (SCM) is also introduced which is an amalgamation of both CCA and SM. Another work incorporating SM, correlation matching (CM), and SCM is presented in [6] where CM is an unsupervised technique that models cross-modal associations.

Various CCA variants have been applied recently in the cross-modal retrieval task. An unsupervised kernel canonical correlation analysis (KCCA) approach has been introduced in [7] which analyzes the relation between image annotation by humans and the respective significance of objects and their contour in the scene. A multi-label KCCA technique is introduced in [8] that augments KCCA with high-level semantic information in multi-label annotations. To overcome the issues in traditional 2-view CCA, [9] has proposed an improved version of that, dubbed, Improved CCA (ICCA). It has also expanded 2-view CCA to 4-view CCA and embed it into a progressive framework for reducing overfitting. In [10], authors have utilized the probabilistic interpretation of CCA. CCA is used for projecting modalities' features to a latent space and probabilistic interpretation is employed for computing the representative distribution of the latent variables for each class. [11] proposed the deep canonical correlation analysis (DCCA) with progressive and hypergraph learning. DCCA is utilized for image and text pair mapping onto a shared latent space and the hypergraph semantic embedding (HSE) approach is used for extracting latent semantics from the text.

2.1.2. *Graph regularization methods*

A graph model is constructed in the process of graph regularization that maintains affinity between the projected data through the edges of the graph model. Graph regularization increases semantic relevance and learns inter and intra-modality similarity. Authors in [12] have proposed a graph regularization and modality dependence (GRMD) approach for making complete utilization of potential correlation among diverse modalities for effective cross-modal retrieval. For individual retrieval task, a separate projection matrix is learned considering semantic and feature correlation. Afterward, the internal structure of the original feature space is used for constructing an adjacent graph having semantic information constraints that makes dissimilar heterogeneous data closer to the related semantic information. An optimal solution for cross-modal retrieval is provided by [13] 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 (SC-CMR) makes use of the semantic information present in unannotated 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 [14]. 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 [15] 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 [16] 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. In [17], authors have proposed a joint graph regularization based modality dependent cross-media retrieval (JGRMDCR) method which considers inter and intra-modality similarity and one-on-one correspondence between diverse modality data pairs.

2.1.3. *Generative adversarial networks*

Recently, cross-modal retrieval task has progressed substantially with the use of generative adversarial networks (GAN). However, joint extraction and utilization of both modality-shared and modality-specific features have not been considered well. So, a modality-specific and shared generative adversarial network (MS^2GAN) approach is proposed in [18] which incorporates two separate sub-networks and a common subnetwork for learning modality-specific and modality-shared features respectively. [19] has introduced a novel end-to-end framework known as adversarial learning based semantic correlation representation (ALSCOR) framework which combines cross-modal representation learning, adversarial, and correlation learning. Non-linear correlation is captured by integrating the CCA model with TxtNet and VisNet representation models. Inspired by zero-shot learning, [20] has proposed a novel ternary adversarial networks with self-supervision (TANSS) technique. The method incorporates three parallel subnetworks: (a) two semantic feature learning subnetworks for preserving modality relationships using semantic features; and (b) a self-supervised semantic subnetwork that supervises semantic feature learning. Adversarial learning is utilized for augmenting the consistency and correlation of the semantic features.

A novel semantic consistent adversarial cross-modal retrieval (SC-ACMR) approach is proposed in [21]. It learns semantic consistent representation for diverse modals under an adversarial learning framework by considering inter and intra-modality semantic similarity. A multi-modal adversarial network (MAN) is proposed in [22] which projects the data onto a shared space where likeness among various modalities can be evaluated using same distance metric. MAN includes a discriminator, multiple modality-specific generators, and a multi-modal discriminant analysis loss. Inspired from the fact that it is difficult to collect large scale multi-modal data, so knowledge in large scale uni-modal data should be fully exploited for enhancing cross-modal retrieval, [23] has proposed a modal-adversarial hybrid transfer network (MHTN). The transition of information from a uni-modal source domain to the cross-modal target domain is realized and cross-modal mutual representation is learned. MHTN comprises of modal-sharing knowledge transfer subnetwork and modal-adversarial semantic learning subnetwork. Former subnetwork jointly transfers information from a huge uni-modal dataset in the source domain to various modalities in the target domain and the latter constructs an adversarial training system between modality discriminator and common representation generator. An adversarial cross-modal retrieval based on dictionary learning (DLA-CMR) framework is introduced in [24]. Adversarial learning extracts the arithmetic features of every modality and dictionary learning aids as feature re-constructor for reconstructing discerning features.

2.1.4. *Deep learning based methods*

The widespread use and benefits of deep neural network (DNN) in single modality retrieval cases have initialized its rapid use in cross-modal retrieval tasks as well. A deep neural network based technique is proposed in [25] which learns common representation for all considered modalities. The technique is known as hybrid representation learning (HRL) in which stacked restricted Boltzmann machines are used for extracting modality-friendly representation and a multi-modal deep belief network is exploited for extracting modality-mutual representation. Shared semantic space with correlation alignment (S^3CA) is introduced in [26] for multi-modal data representation. Non-linear correlations of multi-modal data distributions are aligned in deep neural networks constructed for dissimilar data. In [27], a novel deep adversarial metric learning (DAML) method has been proposed which nonlinearly maps labeled multi-modal data pairs into a common latent feature subspace. DAML augments the inter-class disparity and reduces the intra-class disparity and the variance of each multi-modal data pair of the same class. [28] has presented a multi-modal semantic autoencoder (MMSAE) method for cross-modal retrieval. It consists of a two-stage learning procedure in which multi-modal mappings are learned for projecting multi-modal data onto a low dimensional embedding to preserve feature and semantic information.

In [29], authors 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. A deep semi-supervised cross-modal retrieval framework is proposed in [30] which can effectually tackle both labeled and unlabeled multi-modal data. A label prediction component is utilized in predicting labels for unlabeled training data and a shared representation is learned for the modalities. Various multi-modal datasets suffer from a weak-pairing issue where one concept of data samples in one mode corresponds to the same concept of data samples in another mode rather than direct sample-to-sample correspondence among modalities which introduces a challenge in cross-modal retrieval. Authors in [31] have proposed a novel scalable hierarchical learning framework dubbed deep dictionary learning (DDL) to handle this challenge which considers the cross-modal representations without direct correspondence and minimal concept label supervision. For dealing with label supervision, a shared classifier is introduced across diverse modalities and for modal invariant representation, a multi-modal low-rank model is proposed.

2.1.5. *Other cross-modal retrieval methods*

A multi-modal multi-class boosting framework (MMBoost) is proposed in [32] which can capture both inter-modal semantic correlation and intra-modal semantic information simultaneously. Few researchers take cross-modal retrieval as a learning to rank task. New learning to rank with relational graph and pointwise constraint (LR^2GP) approach is proposed in [33] which aims to optimize the ranking model. In [34], a cross-media framework has been introduced that is based upon linear discriminant analysis. In order to project low-level characteristics into a common feature space by transformation matrices, it incorporates the association between visual and textual features to learn a pair of projection matrices. Hence, a discerning attribute of one mode is transferred to the respective attribute of the other mode using the correlation analysis procedure. A semi-supervised modality dependent cross-modal retrieval approach is introduced in [35] that is based upon coupled feature selection.

A task-dependent and query-dependent subspace learning (TQSL) method has been proposed in [36]. 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, [37] has proposed a correlation-based cross-modal subspace learning model using kernel dependence maximization (KDM). 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 [38] which distinguishes among diverse concepts and utilizes the shared data related to semantic overlap. In [39], authors have presented a semi-supervised modality-dependent cross-media retrieval (SMDCR) method. SMDCR completely utilizes the global data distribution property and semantic data related to both labeled and unlabeled samples.

Thanks to the relentless effort performed by researchers, great advances have been observed in the field of cross-modal retrieval recently. However, most of them have utilized the common subspace learning procedure and the modality features available with the respective datasets. It is required to extract the highly discriminative and non-overlapping modality features for representation which consequently affects the overall efficiency of cross-modal retrieval task. Hence, Zernike moments have been utilized for visual feature representation in the proposed approach and associative learning has also been incorporated in the form of Hebbian learning for integrating two SOMs. Table (1) presents the characteristics of the techniques which are considered for comparison with the proposed technique.

2.2. *Associative memory based techniques*

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 different from the algorithmic methods discussed in the previous sub-section as they 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.

In [40], authors have introduced the use of auto-associative Hopfield network for the cross-modal retrieval process. 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 [41] 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 [42] 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.

Table 1: 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 [5]	U	✓						
SM [5]	S			✓				
SCM [5]	S	✓		✓				
DDL [31]	-	✓		✓		✓		✓
DLA-CMR [24]	-			✓	✓	✓	✓	
DAML [27]	S	✓			✓		✓	✓
SCCMR [13]	Se	✓	✓	✓	✓			
CSGL [14]	S	✓	✓	✓				
SGRCR [15]	S	✓	✓		✓			
CCDCR [16]	S		✓	✓	✓			
CR-CDSL [29]	S	✓		✓				✓
TQSL [36]	S	✓	✓	✓	✓			
KDM [37]	S	✓		✓	✓			
MJSL [38]	S	✓		✓				
SMDCR [39]	Se	✓		✓				

A multiSOM approach has been proposed in [43]. The working of the traditional 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. In [44], authors have proposed an associative self-organizing framework to integrate multi-modal inputs of vision, language and motor programs for producing complex robot behaviors.

In [45], a novel approach has been presented for constructing multi-modal representations by learning a language-to-vision mapping and its result has been used to create multi-modal embeddings. 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 [46]. The issue of learning multi-modal word representations by linking visual, auditory, and textual inputs has been considered.

3. 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 have been considered here for image vector creation because of their efficacy in extracting prominent image features [47, 48, 49]. Latent Dirichlet Allocation (LDA) [50] model is utilized for extracting text features due to its prominence in text analytic area [51, 52].

3.1. 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 and they are briefly defined in the following paragraph. Figure (4) shows the steps followed by each image for ZM extraction which can distinguish it from other images in the data. Each image is preprocessed before calculation of ZM which includes image resizing to 1000 × 1000 size, RGB to grayscale conversion (if it is not grayscale), and image normalization. These steps are followed in a similar way as described in [53].

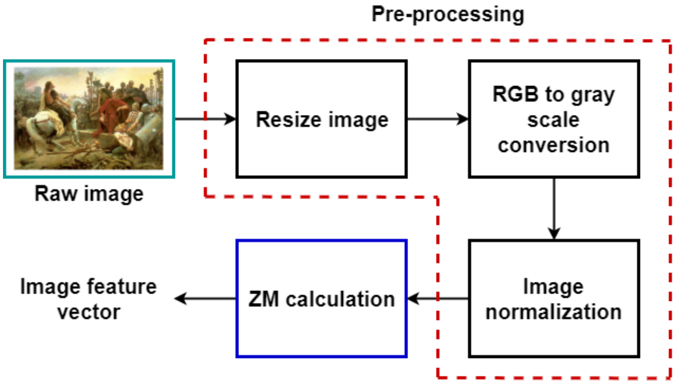


Figure 4: Process followed by each image for ZM extraction

Zernike Moments (ZM): Image moments represent the weighted average of the intensity values of image pixels to obtain the scalar quantities for image interpretation. ZM are a type of continuous orthogonal moments. The benefits of utilizing the orthogonal moments are that they capture the features which are robust to noise, least overlapped or redundant, rotation, scale, and translation invariant [54]. 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* [55]. 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). 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 5); and (2) outer circle mapping (figure 6) [56]. 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 5(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.

According to [57], 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:

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

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

$$V_{st}(r,\mu) = R_{st}(r) e^{it\mu} \tag{2}$$

which satisfies $s \geq 0$, $0 \leq \mid t \mid \leq s$, $s- \mid t \mid =$ even, $i = \sqrt{-1}$ and $\mu = \arctan(y/x)$. (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:

$$R_{st}(r) = \sum_{k=0}^{(s-|t|)/2} (-1)^k \times \frac{(s-k)!}{k!(\frac{s+|t|}{2}-k)!(\frac{s-|t|}{2}-k)!} r^{s-2k} \tag{3}$$

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 [58]. The obtained number of moments (NoM) according to given order s can be evaluated as:

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

3.2. 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 (7) 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.

Latent Dirichlet Allocation (LDA): Topic Modeling is a prominent technique in text mining and detecting relations among textual documents [59]. 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 [50]. 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 (8) 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.

The generative procedure followed for each document $\mathbf{w}$ in a corpus C in LDA is given below [50]:

1. Choose $R \sim$ Poisson(ξ).
2. Choose $\eta \sim$ Dir(γ).
3. For each word w_r in a document, choose:
   (a) a topic $z_r \sim$ Multinomial(η).
   (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

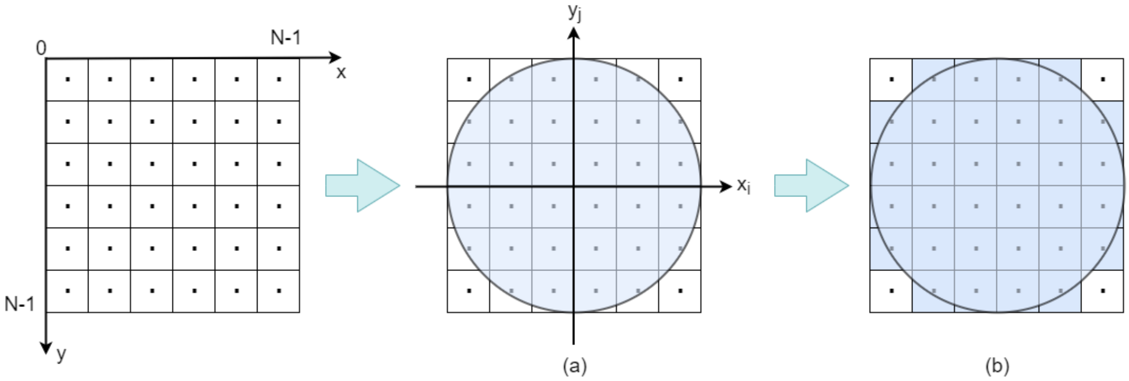


Figure 5: Inner circle mapping technique. (a) image mapping onto unit disk; (b) inscribed disk approximated by square grids

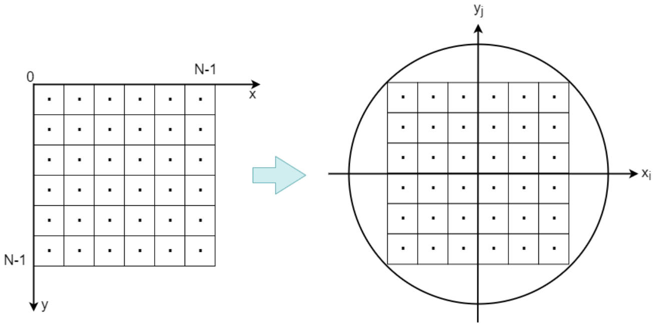


Figure 6: Outer circle mapping technique where complete image is mapped inside the disk

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:

$$p(\eta|\gamma) = \frac{\Gamma(\sum_{i=1}^{k} \gamma_i)}{\prod_{i=1}^{k} \Gamma(\gamma_i)} \eta_1^{\gamma_1 - 1} ... \eta_k^{\gamma_k - 1} \tag{5}$$

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:

$$p(\eta, \mathbf{z}, \mathbf{w}|\gamma, \delta) = p(\eta|\gamma) \prod_{r=1}^{R} p(z_r|\eta) p(w_r|z_r, \delta) \tag{6}$$

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:

$$p(\mathbf{w}|\gamma, \delta) = \int p(\eta|\gamma) \left(\prod_{r=1}^{R} \sum_{z_r} p(z_r|\eta) p(w_r|z_r, \delta) \right) d\eta \tag{7}$$

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

$$p(C|\gamma, \delta) = \prod_{c=1}^{Q} \int p(\eta_c|\gamma) \left(\prod_{r=1}^{R_c} \sum_{z_{cr}} p(z_{cr}|\eta_c) p(w_{cr}|z_{cr}, \delta) \right) d\eta_c \tag{8}$$

4. Proposed Technique

4.1. 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 $N_1 + N_2 = N$.

4.2. Proposed hybrid-SOM based cross-modal retrieval method

4.2.1. Traditional SOM

It is also popular as *Kohonen map* after the name of its inventor *Teuvo Kohonen* who proposed it in 1982 [60]. The funda-

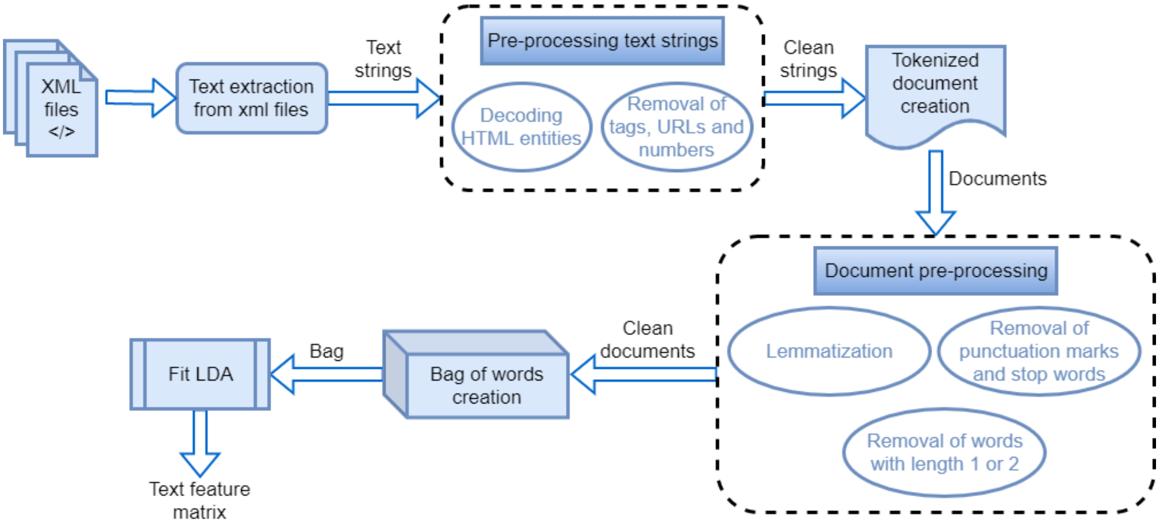

XML
files
</>
Text extraction from xml files
Text strings
Pre-processing text strings
Decoding HTML entities
Removal of tags, URLs and numbers
Clean strings
Tokenized document creation
Documents
Document pre-processing
Lemmatization
Removal of punctuation marks and stop words
Removal of words with length 1 or 2
Clean documents
Bag of words creation
Bag
Fit LDA
Text feature matrix

Figure 7: Process flow for text feature matrix creation

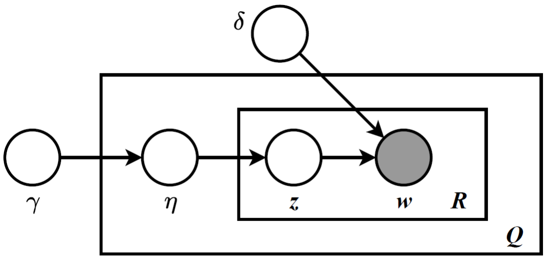


Figure 8: Graphical model representation of LDA

mental 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 [61]. 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 [62]. Figure (9) 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 [63]. Based upon the BMU, the weights of the neighboring neurons are also modified.

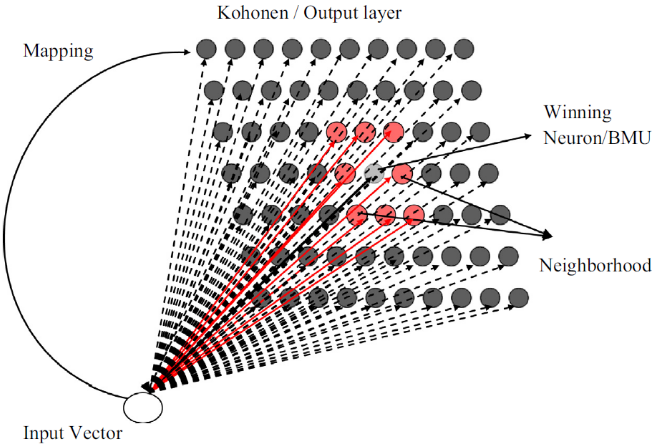


Figure 9: Representation of traditional SOM [63]

Table 2: 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 (2) presents the variable notations and definitions which are being utilized in SOM learning algorithm. The procedure followed in traditional SOM learning is as follows [62]:

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.

$$\|x_i - w_c\| = \min_k \|x_i - w_k\| \tag{9}$$

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.

$$w_j(t+1) = w_j(t) + \Delta w_j(t) \tag{10}$$

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

$$\Delta w_j(t) = \alpha(t) n_{cj}(t)(x_i(t) - w_j(t)) \tag{11}$$

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:

$$n_{cj}(t) = exp\left(-\frac{\|r_c - r_j\|^2}{2\sigma^2(t)}\right), \tag{12}$$

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 9). 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.

4.2.2. 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 [64]. 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. 13).

$$\Delta w_{ij} = \alpha \times x_i \times y_j \tag{13}$$

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. 14):

$$w_{ij}(n) = w_{ij}(n-1) + \Delta w_{ij} \tag{14}$$

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 (10) 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 $m \times n \times p \times q$.

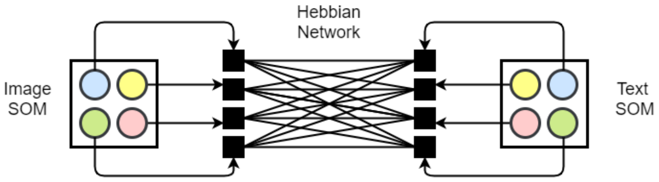


Figure 10: Architecture of two SOMs (image and text) connected using Hebbian network

For the implementation of the proposed technique, features are extracted from the available images and corresponding text as mentioned in section (3). 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 (15) and the Hebbian 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.

$$hebbLink(classes_I(i), classes_T(i)) += LR*whebb_I(i)*whebb_T(i) \tag{15}$$

where $1 \leq i \leq length(classes_I)$ as the hebbian network is trained for the total number of inputs [65]. Here, *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 (12) 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. Although, HSOM is a supervised approach individually as the trained SOM nodes are acting as labels for the input instances for hybrid model training. However, we are calling the overall algorithm as unsupervised as there is no requirement of class information for each of the inputs in the beginning. The algorithms (1,2) present all the steps followed for the implementation of the proposed technique. Figure (11) 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.

5. Experimental analysis

5.1. Dataset

The proposed approach has been tested on *Wikipedia*[1] [5] 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 (3).

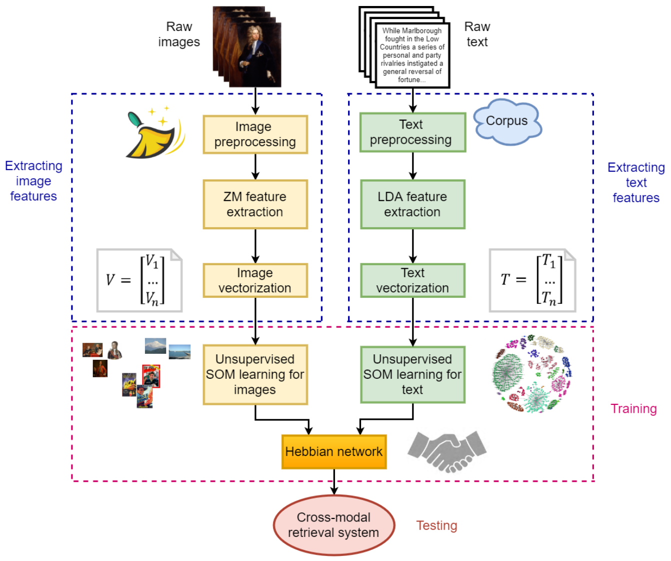

Raw images
Raw text
While Marlborough fought in the Low Countries a series of personal and party rivalries instigated a general reversal of fortune...
Extracting image features
Image preprocessing
ZM feature extraction
Image vectorization
$V = \begin{bmatrix} V_1 \\ \dots \\ V_n \end{bmatrix}$
Text preprocessing
Corpus
LDA feature extraction
Text vectorization
$T = \begin{bmatrix} T_1 \\ \dots \\ T_n \end{bmatrix}$
Extracting text features
Unsupervised SOM learning for images
Unsupervised SOM learning for text
Training
Hebbian network
Cross-modal retrieval system
Testing

Figure 11: Process flow of image and text training for the proposed hybrid cross-modal retrieval system

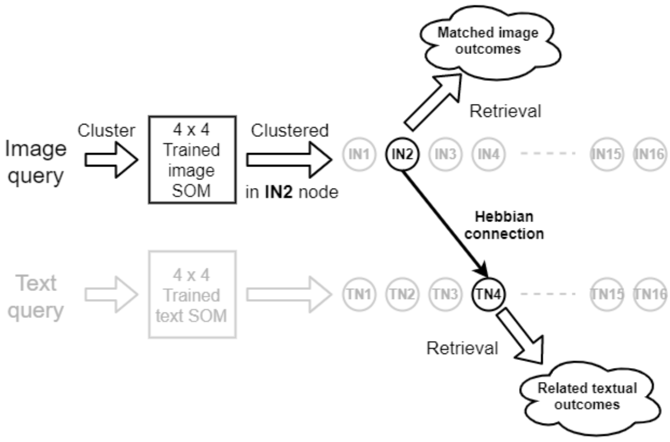


Figure 12: Handling of a test query. The dark portion of the figure depicts the procedure followed by an image query. Both image and text modalities are retrieved in the end.

5.2. 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) [66]. 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:

$$AP = \frac{1}{R} \sum_{y=1}^{Y} P(y) rel(y) \tag{16}$$

where R represents the ground truth positives or the number of relevant results in the retrieved results [67], $P(y)$ depicts the precision of top y retrieved results, and the value of Y is different for each of the test (image/text) instances because the number of retrieved outcomes are different for each test query. If

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

Algorithm 1 Algorithm of proposed technique 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 $1000 \times 1000$
    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
    $net_I \leftarrow selforgmap([dimension1dimension2], 200)$    ▷ Configure image SOM
    $net_T \leftarrow selforgmap([dimension1dimension2], 200)$    ▷ 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 (15)
    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
procedure CREATION OF $Anet_I$
    $net_I Size, net_T Size \leftarrow dimension1 \times dimension2$    ▷ Size of image and text net ($net_I$, $net_T$) in Hebbian network
    for $i \leftarrow 1$ to $net_I Size$ do
        $maxtemp \leftarrow 0, maxindex \leftarrow -1$    ▷ Initializing the temporary variables
        for $j \leftarrow 1$ to $net_T Size$ do
            if $hebbLink(i, j) > maxtemp$ then    ▷ Checking for the maximum hebbLink weight
                $maxtemp = hebbLink(i, j)$
                $maxindex = j$
            end if
        end for
        $Anet_I(i) = maxindex$
    end for
end procedure
procedure CREATION OF $Anet_T$
    for $i \leftarrow 1$ to $net_T Size$ do
        $maxtemp \leftarrow 0, maxindex \leftarrow -1$    ▷ Initializing the temporary variables
        for $j \leftarrow 1$ to $net_I Size$ do
            if $hebbLink(j, i) > maxtemp$ then    ▷ Checking for the maximum hebbLink weight
                $maxtemp = hebbLink(j, i)$
                $maxindex = j$
            end if
        end for
        $Anet_T(i) = maxindex$
    end for
end procedure
```

Table 3: 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

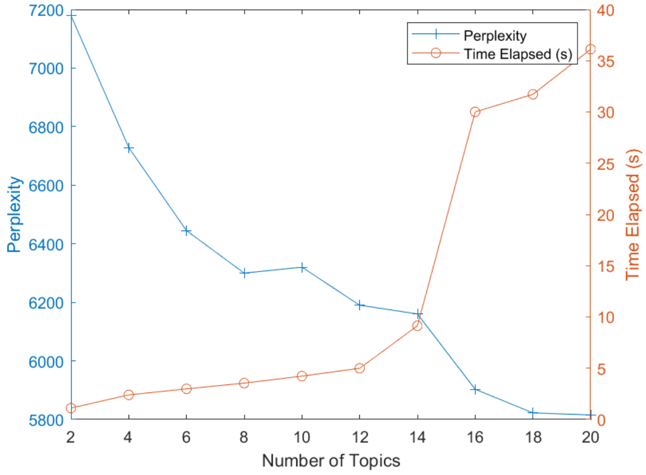


Figure 13: Perplexity and time analysis for choosing an appropriate number of topics for the LDA model

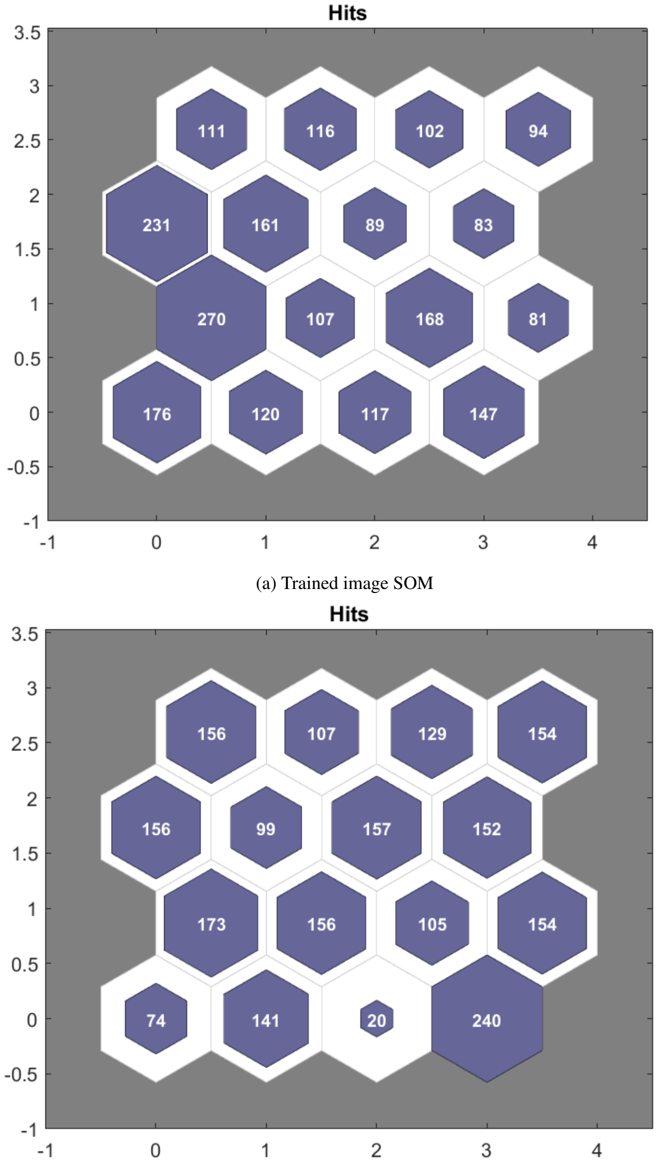
the y^{th} retrieved result is relevant then $rel(y) = 1$ and otherwise 0. Now, MAP can be calculated as:

$$MAP = \frac{1}{N} \sum_{n=1}^{N} AP \tag{17}$$

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

5.3. Comparison methods

1. *CCA* [5] is a fundamental subspace learning based method. It finds the pair of projections for different modes so that the relation between them is augmented.



(a) Trained image SOM

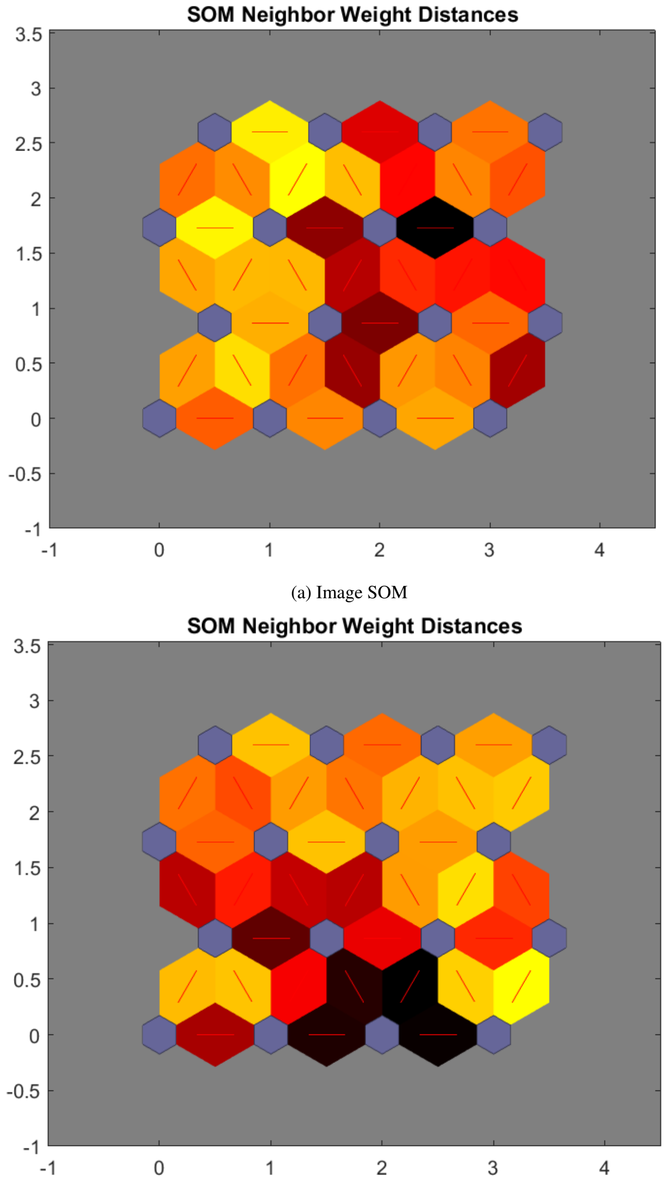
(b) Trained text SOM

Figure 14: Input train data distribution after individual SOM training

2. *SM* [5] 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* [5] 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* [31] 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.



(a) Image SOM

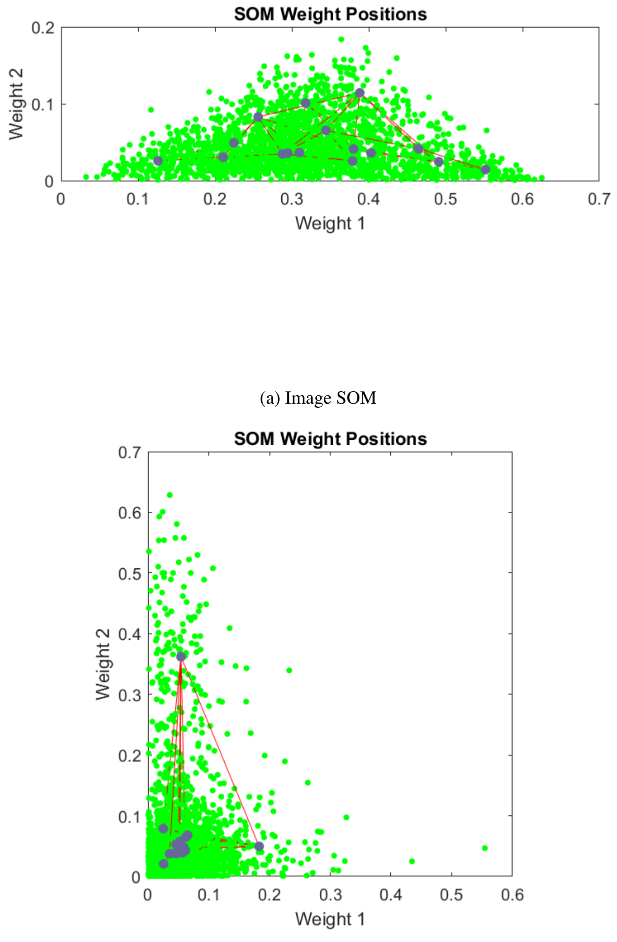
(b) Text SOM

Figure 15: Neighbor Distances among respective SOM nodes. Darker shade denotes larger distance.

5. *DLA-CMR* [24] 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 re-constructor 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* [27] 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



(a) Image SOM

(b) Text SOM

Figure 16: Location of data points and weight vectors

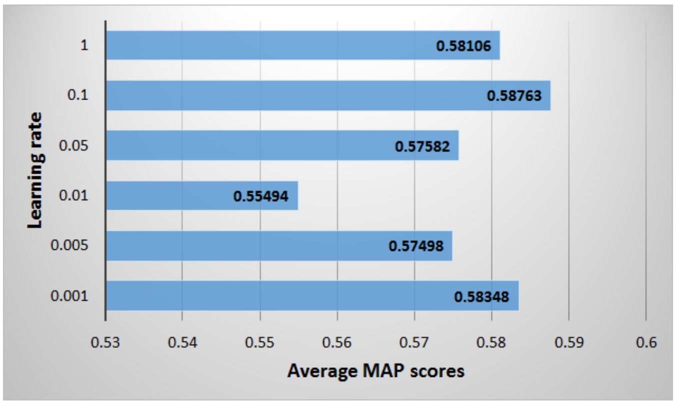


Figure 17: Average MAP scores at different values of learning rate for training Hebbian network

regularization is added by the introduction of adversarial learning.

7. *SCCMR* [13] 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 in the potential subspace of paired text and images and text and images with identical semantics.

8. *CSGL* [14]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* [15] 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* [16] 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* [29] 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* [36] 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* [37] is a cross-modal subspace learning framework which is based upon correlation. Unlike most other methods that directly maximize feature correlations across

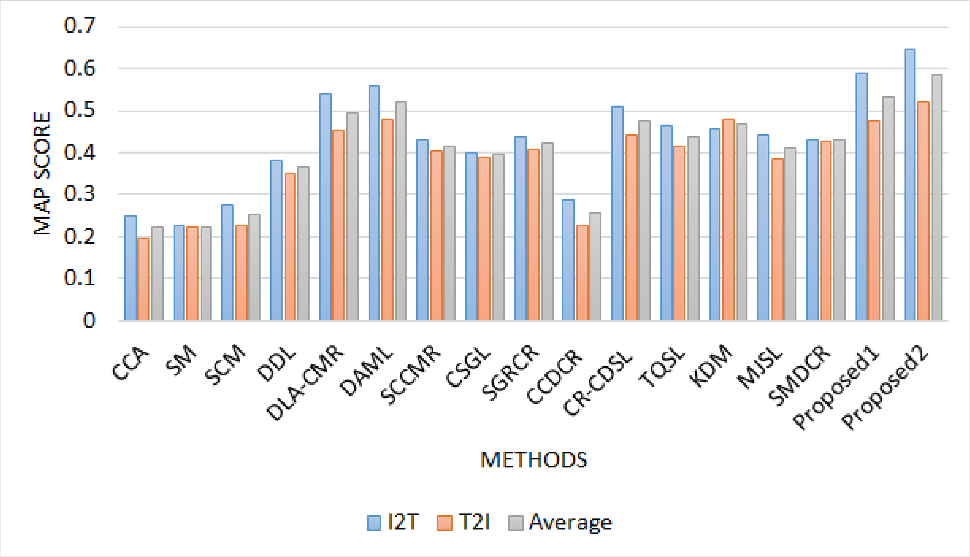


Figure 18: Performance analysis based on MAP scores

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

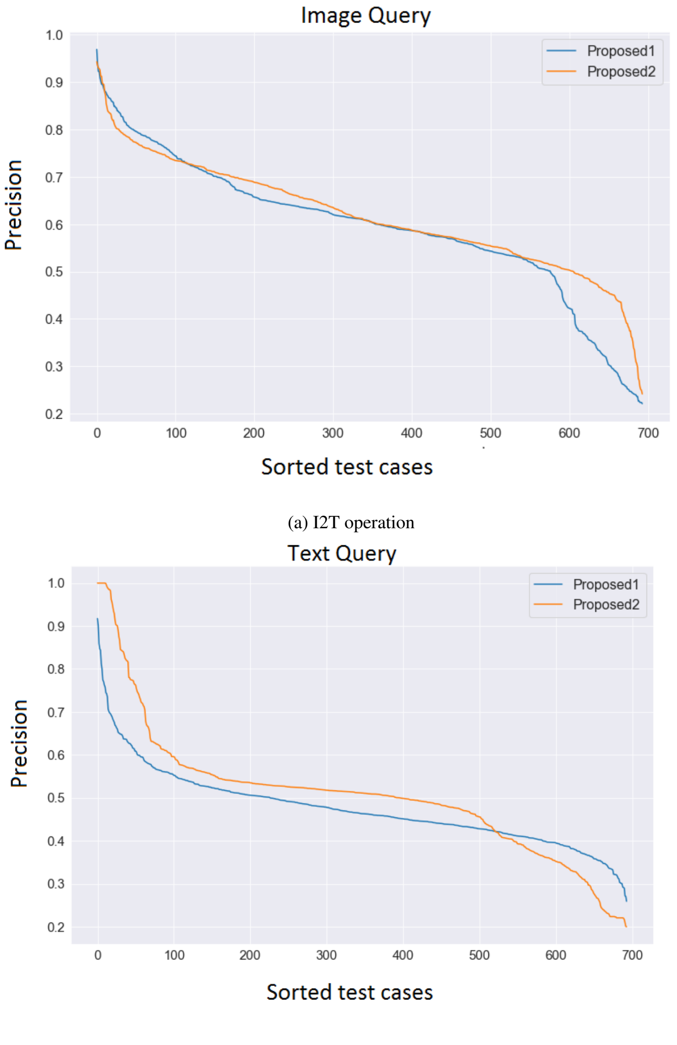
15. *SMDCR* [39] is a semi-supervised cross-modal technique which is modality-dependent and uses both labeled and unlabeled samples for getting two couple of projection matrices. It utilizes the feature distance for representing the semantic knowledge of unlabeled samples in the optimization process for getting full use of data structural information. It fully utilizes the semantic knowledge of whole multi-modal data and data distribution property.

5.4. 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 (13). 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.

5.5. Model training

Figure (14) 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 (15) 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 15a). 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 15b) 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 (14b), 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 (16). 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 (17), it is noticeable that the 0.1 learning rate value is suitable for the experimental analysis.



(a) I2T operation

(b) T2I operation

Figure 19: Curves depicting the sorted precision values for test queries (693 in our case)

5.6. Results

Table (4) 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 [6] for representing images and collateral text. These features are freely provided by the authors on the *link*[2] 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 training image[3] 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[4], are fit by LDA model [50] by utilizing the implementation of [68]. 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,

Table 4: 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 [5]	0.249	0.196	0.223	H
SM [5]	0.225	0.223	0.224	H
SCM [5]	0.277	0.226	0.252	H
DDL [31]	0.2832	0.2615	0.2724	H
	0.3812	0.3501	0.3657	D
DLA-CMR [24]	0.369	0.261	0.315	H
	0.539	0.453	0.496	D
DAML [27]	0.356	0.267	0.322	H
	0.559	0.481	0.52	D
SCCMR [13]	0.431	0.403	0.417	D
CSGL [14]	0.3996	0.3904	0.395	H
SGRCR [15]	0.284	0.227	0.2555	H
	0.4365	0.406	0.421	D
CCDCR [16]	0.2849	0.2253	0.2551	H
CR-CDSL [29]	0.348	0.249	0.299	H
	0.508	0.442	0.475	D
TQSL [36]	0.463	0.415	0.439	D
KDM [37]	0.4562	0.4785	0.4674	D
MJSL [38]	0.4432	0.3832	0.4132	D
SMDCR [39]	0.284	0.232	0.258	H
	0.43	0.428	0.429	D
Proposed1	0.5872	0.4744	0.5308	H
Proposed2	**0.6461**	**0.5228**	**0.5844**	-

Table 5: 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

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

[3] https://lear.inrialpes.fr/people/dorko/downloads.html

[4] http://www.nltk.org/

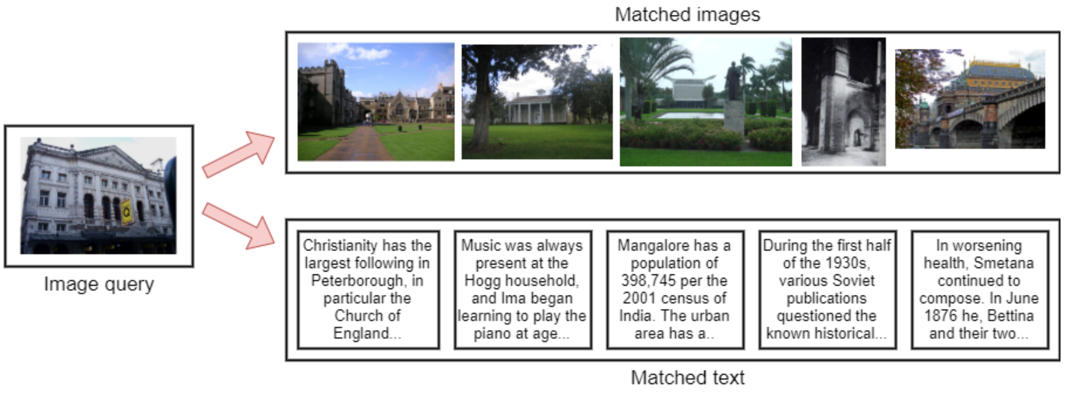

Matched images
Image query
Christianity has the largest following in Peterborough, in particular the Church of England...
Music was always present at the Hogg household, and Ima began learning to play the piano at age...
Mangalore has a population of 398,745 per the 2001 census of India. The urban area has a..
During the first half of the 1930s, various Soviet publications questioned the known historical...
In worsening health, Smetana continued to compose. In June 1876 he, Bettina and their two...
Matched text

Figure 20: Retrieved image and text results using an image query

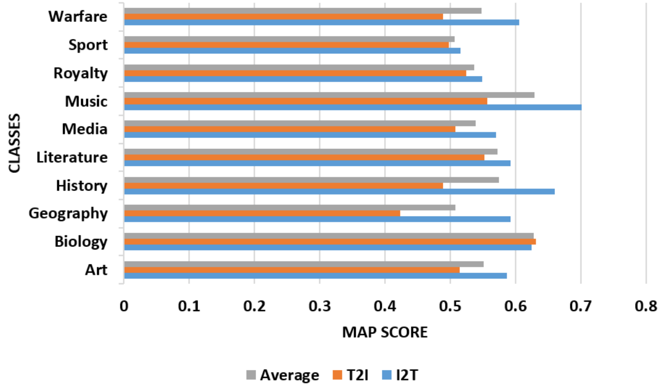


Figure 21: Performance chart based on MAP scores for each class

they have been considered in this study also for comparative analysis.

In the table, *Proposed1* represents the results obtained using handcrafted features, and *Proposed2* depicts the results achieved using 12-d ZM for images and 14-d LDA for text. It is evident from the table that the Proposed2 approach is better than the other methods and so the MAP scores are highlighted. Figure (18) shows the performance chart of all the methods based on their MAP scores and figure (21) presents the class-wise performance of the proposed technique. Table (5) shows the I2T, T2I and their average MAP score values for respective dataset categories. Figure (19) 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 (20) illustrates a few matched images and text results retrieved using an image query on trained Proposed2 model.

5.6.1. *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 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) 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 than a single source (such as only text or image). Figure (22) is similar to figure (19) 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 (22a), however, figure (22b) 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.

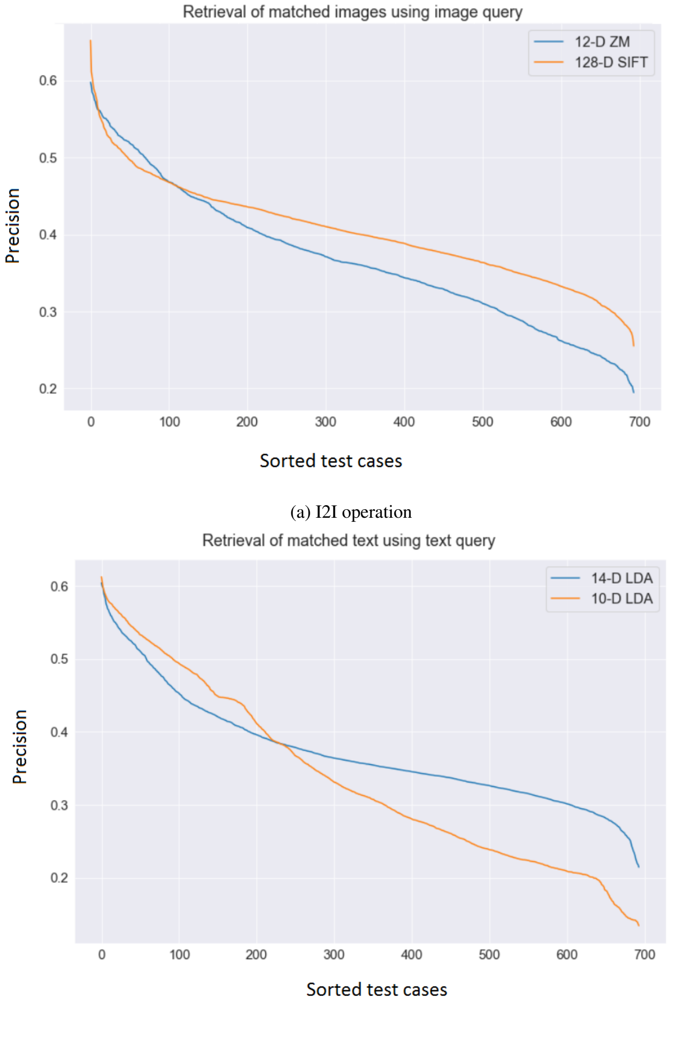
Table 6: Comparison of MAP scores obtained using traditional SOM and HSOM

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**

5.6.2. *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:

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 [49].



(a) I2I operation

(b) T2T operation

Figure 22: Curves demonstrating the sorted precision values for test queries in case of uni-modal retrieval

2. LDA features that are used for text representation provide well-defined inference procedures for even unseen documents [50]. 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 [60]. Moreover, SOM has shown its effectiveness recently in various application areas such as speech recognition [69], mental stress detection [70], coronary heart disease diagnosis [71], and for extracting features as an add-on for better network intrusion detection [72].

4. SOM helps in easy interpretation and understanding of data [60]. Similarities in data can be easily observed and visualized using SOM. It has the ability to cluster even large or complex datasets [61].

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 [64]. 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 [73], 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 [74] 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 [75].

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 [76].

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 [77].

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 [78]. 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. Conclusion and future scope

This paper 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 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 and results prove the efficacy of the proposed technique in the field of cross-modal retrieval.

Image and text SOM grid size is a parameter for subjective tuning. Although, the results obtained using the proposed approach are promising, however, image semantics are required to be considered more carefully for better performance. In the future, diverse image and textual noise removal techniques should be considered for further improvement in the MAP scores. The presented framework can be utilized in the medical field by utilizing suitable feature extractors based on the images. Different associative learning techniques such as the Hopfield network and Bi-directional auto-associative memory network can also be incorporated into this study.

Declaration of Competing Interest

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

References

[1] B. E. Stein, M. A. Meredith, The merging of the senses., The MIT Press, 1993.

[2] B. E. Stein, M. A. Meredith, W. S. Huneycutt, L. McDade, Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli, Journal of Cognitive Neuroscience 1 (1) (1989) 12–24.

[3] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in statistics, Springer, 1992, pp. 162–190.

[4] C. Guo, D. Wu, Canonical correlation analysis (cca) based multi-view learning: An overview, arXiv preprint arXiv:1907.01693 (2019).

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

[6] J. C. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. R. Lanckriet, R. Levy, N. Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE transactions on pattern analysis and machine intelligence 36 (3) (2013) 521–535.

[7] S. Hwang, K. Grauman, Learning the relative importance of objects from tagged images for retrieval and cross-modal search, International journal of computer vision 100 (2) (2012) 134–153.

[8] Y. Jia, L. Bai, S. Liu, P. Wang, J. Guo, Y. Xie, Semantically-enhanced kernel canonical correlation analysis: a multi-label cross-modal retrieval, Multimedia Tools and Applications 78 (10) (2019) 13169–13188.

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

[10] M. Katsurai, T. Ogawa, M. Haseyama, A cross-modal approach for extracting semantic relationships between concepts using tagged images, IEEE transactions on multimedia 16 (4) (2014) 1059–1074.

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

[12] G. Wang, H. Ji, D. Kong, N. Zhang, Modality-dependent cross-modal retrieval based on graph regularization, Mobile Information Systems 2020 (2020).

[13] G. Xu, X. Li, Z. Zhang, Semantic consistency cross-modal retrieval with semi-supervised graph regularization, IEEE Access 8 (2020) 14278–14288.

[14] G. Xu, X. Li, L. Shi, Z. Zhang, A. Zhai, Combination subspace graph learning for cross-modal retrieval, Alexandria Engineering Journal (2020).

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

[16] M. Zhang, H. Zhang, J. Li, Y. Fang, L. Wang, F. Shang, Multi-modal graph regularization based class center discriminant analysis for cross modal retrieval, Multimedia Tools and Applications 78 (19) (2019) 28285–28307.

[17] J. Yan, H. Zhang, J. Sun, Q. Wang, P. Guo, L. Meng, W. Wan, X. Dong, Joint graph regularization based modality-dependent cross-media retrieval, Multimedia Tools and Applications 77 (3) (2018) 3009–3027.

[18] F. Wu, X.-Y. Jing, Z. Wu, Y. Ji, X. Dong, X. Luo, Q. Huang, R. Wang, Modality-specific and shared generative adversarial network for cross-modal retrieval, Pattern Recognition (2020) 107335.

[19] L. Zhu, J. Song, X. Wei, L. Jun, Adversarial learning based semantic correlation representation for cross-modal retrieval (2020).

[20] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, X. Li, Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval, IEEE transactions on cybernetics 50 (6) (2019) 2400–2413.

[21] W. Ou, R. Xuan, J. Gou, Q. Zhou, Y. Cao, Semantic consistent adversarial cross-modal retrieval exploiting semantic similarity, Multimedia Tools and Applications (2019) 1–18.

[22] P. Hu, D. Peng, X. Wang, Y. Xiang, Multimodal adversarial network for cross-modal retrieval, Knowledge-Based Systems 180 (2019) 38–50.

[23] X. Huang, Y. Peng, M. Yuan, Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval, IEEE transactions on cybernetics (2018).

[24] F. Shang, H. Zhang, L. Zhu, J. Sun, Adversarial cross-modal retrieval based on dictionary learning, Neurocomputing 355 (2019) 93–104.

[25] W. Cao, Q. Lin, Z. He, Z. He, Hybrid representation learning for cross-modal retrieval, Neurocomputing 345 (2019) 45–57.

[26] Z. Yang, Z. Lin, P. Kang, J. Lv, Q. Li, W. Liu, Learning shared semantic space with correlation alignment for cross-modal event retrieval, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 16 (1) (2020) 1–22.

[27] X. Xu, L. He, H. Lu, L. Gao, Y. Ji, Deep adversarial metric learning for cross-modal retrieval, World Wide Web 22 (2) (2019) 657–672.

[28] Y. Wu, S. Wang, Q. Huang, Multi-modal semantic autoencoder for cross-modal retrieval, Neurocomputing 331 (2019) 165–175.

[29] B. Zhang, L. Zhu, J. Sun, H. Zhang, Cross-media retrieval with collective deep semantic learning, Multimedia Tools and Applications 77 (17) (2018) 22247–22266.

[30] D. Mandal, P. Rao, S. Biswas, Semi-supervised cross-modal retrieval with label prediction, IEEE Transactions on Multimedia (2019).

[31] H. Liu, F. Wang, X. Zhang, F. Sun, Weakly-paired deep dictionary learning for cross-modal retrieval, Pattern Recognition Letters 130 (2020) 199–206.

[32] S. Wang, Z. Dou, D. Chen, H. Yu, Y. Li, P. Pan, Multimodal multiclass boosting and its application to cross-modal retrieval, Neurocomputing 357 (2019) 11–23.

[33] Q. Xu, M. Li, M. Yu, Learning to rank with relational graph and pointwise constraint for cross-modal retrieval, Soft Computing 23 (19) (2019) 9413–9427.

[34] Y. Qi, H. Zhang, B. Zhang, L. Wang, S. Zheng, Cross-media retrieval based on linear discriminant analysis, Multimedia Tools and Applications 78 (17) (2019) 24249–24268.

[35] E. Yu, J. Sun, L. Wang, W. Wan, H. Zhang, Coupled feature selection based semi-supervised modality-dependent cross-modal retrieval, Multimedia Tools and Applications 78 (20) (2019) 28931–28951.

[36] L. Wang, L. Zhu, E. Yu, J. Sun, H. Zhang, Task-dependent and query-dependent subspace learning for cross-modal retrieval, IEEE Access 6 (2018) 27091–27102.

[37] M. Xu, Z. Zhu, Y. Zhao, F. Sun, Subspace learning by kernel dependence maximization for cross-modal retrieval, Neurocomputing 309 (2018) 94–105.

[38] E. Yu, J. Li, L. Wang, J. Zhang, W. Wan, J. Sun, Multi-class joint subspace learning for cross-modal retrieval, Pattern Recognition Letters 130 (2020) 165–173.

[39] X. Dong, J. Sun, P. Duan, L. Meng, Y. Tan, W. Wan, H. Wu, B. Zhang, H. Zhang, Semi-supervised modality-dependent cross-media retrieval, Multimedia Tools and Applications 77 (3) (2018) 3579–3595.

[40] R. Shriwas, P. Joshi, V. M. Ladwani, V. Ramasubramanian, Multi-modal associative storage and retrieval using hopfield auto-associative memory network, in: International Conference on Artificial Neural Networks, Springer, 2019, pp. 57–75.

[41] J.-W. Ha, B.-H. Kim, B. Lee, 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, Springer, 2010, pp. 76–87.

[42] J.-W. Ha, B.-J. Lee, 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), IEEE, 2012, pp. 3245–3250.

[43] Z. Liu, 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), IEEE, 2014, pp. 1–5.

[44] S. Wermter, C. Weber, M. Elshaw, Associative neural models for biomimetic multi-modal learning in a mirror neuron-based robot, in: Modeling language, cognition and action, World Scientific, 2005, pp. 31–46.

[45] G. Collell, T. Zhang, M.-F. Moens, Learning to predict: A fast reconstructive method to generate multimodal embeddings, arXiv preprint arXiv:1703.08737 (2017).

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

[47] A. J. Fredo, R. Abilash, R. Femi, A. Mythili, C. S. Kumar, Classification of damages in composite images using zernike moments and support vector machines, Composites Part B: Engineering 168 (2019) 77–86.

[48] B. Kaur, S. Singh, J. Kumar, Iris recognition using zernike moments and polar harmonic transforms, Arabian Journal for Science and Engineering 43 (12) (2018) 7209–7218.

[49] H. Aggarwal, D. K. Vishwakarma, Covariate conscious approach for gait recognition based upon zernike moment invariants, IEEE Transactions on Cognitive and Developmental Systems 10 (2) (2017) 397–407.

[50] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.

[51] L. Zheng, Z. Caiming, C. Caixian, Mmdf-lda: An improved multi-modal latent dirichlet allocation model for social image annotation, Expert Systems with Applications 104 (2018) 168–184.

[52] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, T. Zhu, Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter, PloS one 15 (9) (2020) e0239441.

[53] P. Kaur, H. S. Pannu, A. K. Malhi, Plant disease recognition using fractional-order zernike moments and svm classifier, Neural Computing and Applications 31 (12) (2019) 8749–8768.

[54] P. Kaur, H. S. Pannu, A. K. Malhi, Comprehensive study of continuous orthogonal moments—a systematic review, ACM Computing Surveys (CSUR) 52 (4) (2019) 1–30.

[55] Z. von F, Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode, physica 1 (7-12) (1934) 689–704.

[56] A. Aggarwal, C. Singh, Zernike moments-based gurumukhi character recognition, Applied Artificial Intelligence 30 (5) (2016) 429–444.

[57] M. R. Teague, Image analysis via the general theory of moments∗, J. Opt. Soc. Am. 70 (8) (1980) 920–930. doi:10.1364/JOSA.70.000920.

[58] A. Khotanzad, Y. H. Hong, Invariant image recognition by zernike moments, IEEE Transactions on pattern analysis and machine intelligence 12 (5) (1990) 489–497.

[59] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, Multimedia Tools and Applications 78 (11) (2019) 15169–15211.

[60] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological cybernetics 43 (1) (1982) 59–69.

[61] T. Kohonen, Essentials of the self-organizing map, Neural networks 37 (2013) 52–65.

[62] M. Pacella, A. Grieco, M. Blaco, On the use of self-organizing map for text clustering in engineering change process analysis: a case study, Computational intelligence and neuroscience 2016 (2016).

[63] T. Nanda, B. Sahoo, C. Chatterjee, Enhancing the applicability of kohonen self-organizing map (ksom) estimator for gap-filling in hydrometeorological timeseries data, Journal of Hydrology 549 (2017) 133–147.

[64] D. O. Hebb, The organization of behavior: A neuropsychological theory, Psychology Press, 2005.

[65] A. Abraham, Artificial neural networks, Handbook of measuring system design (2005).

[66] Y. Wang, F. Wu, J. Song, X. Li, Y. Zhuang, Multi-modal mutual topic reinforce modeling for cross-media retrieval, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 307–316.

[67] L. Xie, L. Zhu, G. Chen, Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval, Multimedia Tools and Applications 75 (15) (2016) 9185–9204.

[68] G. Doyle, C. Elkan, Accounting for burstiness in topic models, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 281–288.

[69] S. Lokesh, P. M. Kumar, M. R. Devi, P. Parthasarathy, C. Gokulnath, An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map, Neural Computing and Applications 31 (5) (2019) 1521–1531.

[70] J. Tervonen, S. Puttonen, M. J. Sillanpää, L. Hopsu, Z. Homorodi, J. Keränen, J. Pajukanta, A. Tolonen, A. Lämsä, J. Mäntyjärvi, Personalized mental stress detection with self-organizing map: From laboratory to the field, Computers in Biology and Medicine 124 (2020) 103935.

[71] M. Nilashi, H. Ahmadi, A. A. Manaf, T. A. Rashid, S. Samad, L. Shahmoradi, N. Aljojo, E. Akbari, Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates, International Journal of Fuzzy Systems 22 (4) (2020) 1376–1388.

[72] Y. Chen, N. Ashizawa, C. K. Yeo, N. Yanai, S. Yean, Multi-scale self-organizing map assisted deep autoencoding gaussian mixture model for unsupervised intrusion detection, Knowledge-Based Systems 224 (2021) 107086.

[73] G. Amato, F. Carrara, F. Falchi, C. Gennaro, G. Lagani, Hebbian learning meets deep convolutional neural networks, in: International Conference on Image Analysis and Processing, Springer, 2019, pp. 324–334.

[74] S. Bekhouche, F. Dornaika, A. Benlamoudi, A. Ouafi, A. Taleb-Ahmed, A comparative study of human facial age estimation: handcrafted features vs. deep features, Multimedia Tools and Applications 79 (35) (2020) 26605–26622.

[75] S. A. Siddiqui, A. Salman, M. I. Malik, F. Shafait, A. Mian, M. R. Shortis, 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 75 (1) (2018) 374–389.

[76] O. Ghorbanzadeh, T. Blaschke, K. Gholamnia, S. R. Meena, D. Tiede, J. Aryal, Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection, Remote Sensing 11 (2) (2019) 196.

[77] H. S. Pannu, S. Ahuja, N. Dang, S. Soni, A. K. Malhi, Deep learning based image classification for intestinal hemorrhage, Multimedia Tools and Applications 79 (2020) 21941–21966.

[78] P. Kaur, H. S. Pannu, A. K. Malhi, Comparative analysis on cross-modal information retrieval: a review, Computer Science Review 39 (2021) 100336.