The Visual Computer
https://doi.org/10.1007/s00371-023-03126-z

ORIGINAL ARTICLE

Annotate and retrieve in vivo images using hybrid self-organizing map

Parminder Kaur[1,2] · Avleen Malhi[2,3] · Husanbir Pannu[1,2]

Accepted: 23 September 2023

Abstract
Multimodal retrieval has gained much attention lately due to its effectiveness over uni-modal retrieval. For instance, visual features often under-constrain the description of an image in content-based retrieval; however, another modality, such as collateral text, can be introduced to abridge the semantic gap and make the retrieval process more efficient. This article proposes the application of cross-modal fusion and retrieval on real in vivo gastrointestinal images and linguistic cues, as the visual features alone are insufficient for image description and to assist gastroenterologists. So, a cross-modal information retrieval approach has been proposed to retrieve related images given text and vice versa while handling the heterogeneity gap issue among the modalities. The technique comprises two stages: (1) individual modality feature learning; and (2) fusion of two trained networks. In the first stage, two self-organizing maps (SOMs) are trained separately using images and texts, which are clustered in the respective SOMs based on their similarity. In the second (fusion) stage, the trained SOMs are integrated using an associative network to enable cross-modal retrieval. The underlying learning techniques of the associative network include Hebbian learning and Oja learning (Improved Hebbian learning). The introduced framework can annotate images with keywords and illustrate keywords with images, and it can also be extended to incorporate more diverse modalities. Extensive experimentation has been performed on real gastrointestinal images obtained from a known gastroenterologist that have collateral keywords with each image. The obtained results proved the efficacy of the algorithm and its significance in aiding gastroenterologists in quick and pertinent decision making.

Keywords Self-organizing map · Hebbian learning · Oja rule · Cross-modal retrieval · Gastrointestinal endoscopy

1 Introduction

Annotation of images, and illustration of texts, is a skilled task carried out by experienced professionals. Keywords in text and exemplar image features are used simultaneously to compensate and enhance information in one modality (say, visual features) by using the information in another (e.g., keywords). Given the explosive growth of visual information, partly due to the expansion of the Web and partly due to the introduction of sophisticated and inexpensive image capture systems, there is an urgent need to develop programs that can learn to annotate. Automatic annotation systems are among the key areas of research and development nowadays and beyond, and machine learning is the vital technology in developing such systems [1, 2]. Moreover, cross-modal systems have miscellaneous real-life applications such as face-voice matching and retrieval, disaster and emergency management, spoken-to-sign language transcription, emotion recognition, YouTube video categorization, and biomedical image retrieval for helping physicians to visualize similar cases [3].

The study of multimodal communication is being pursued in neurobiology [4], with emphasis on how communication in one modality (say, speech) may complement or suppress information in another mode (for example, vision) in a systematic way [5]. It has been argued that learning to communicate in a multimodal environment is facilitated by processes akin to self-organization in the infant's brain [6].

✉ Parminder Kaur
pkaur60_phd18@thapar.edu

Avleen Malhi
amalhi@bournemouth.ac.uk

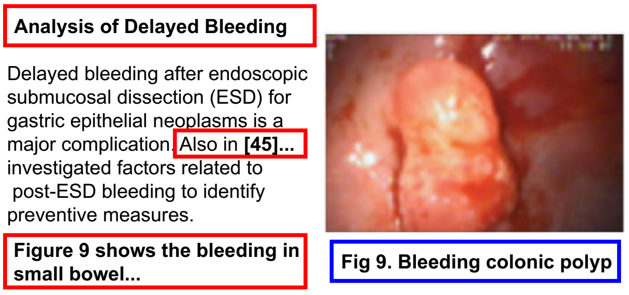
Husanbir Pannu
hspannu@thapar.edu

[1] CSED, Thapar Institute of Engineering and Technology, Patiala, India

[2] Department of Computer Science, Durham University, Durham, UK

[3] Department of Computing and Informatics, Bournemouth University, Poole, UK

Published online: 31 October 2023



Analysis of Delayed Bleeding

Delayed bleeding after endoscopic submucosal dissection (ESD) for gastric epithelial neoplasms is a major complication. Also in [45]... investigated factors related to post-ESD bleeding to identify preventive measures.

Figure 9 shows the bleeding in small bowel...

Fig 9. Bleeding colonic polyp

Fig. 1 Image and collateral text including caption, figure reference, section title, and the related reference for a gastral in vivo image

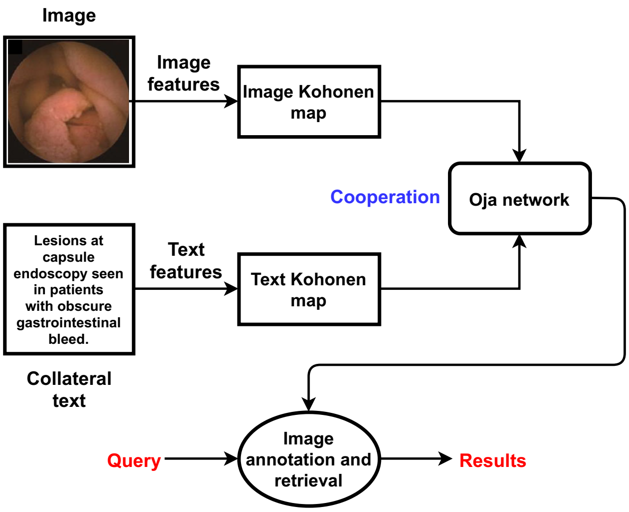


Fig. 2 Flow diagram of the proposed image annotation and retrieval system. Each neural network is validated for its performance and consistency after training before going to the Oja network. SOM is also known as the Kohonen map

Taking the analogy of an animal's brain, where there are identifiable areas for processing a specific mode (such as vision, speech, and touch) that nevertheless communicates with others, one can argue that the brain is a multi-net system with sophisticated communications across specialist areas. Image features alone might be insufficient to discern the dataset using a single modality. For instance, when a layman sees an endoscopy picture, he may not be able to understand it. Perhaps, he can try to understand it with the help of collateral text (such as caption, figure reference, and related citation) describing the image (Fig. 1). So, automatic image annotation, especially using linguistic descriptions of the contents of an image (collection), attempts to bridge the semantic 'gap' between low-level image features and meaning related to objects and events depicted in an image [7].

So, motivated by the fact that multiple modalities provide more information and better understanding, a novel framework has been proposed in this article that can learn to associate different data modalities, for instance, images with keywords, in this study. Two SOM networks are trained with information about an object in two different modalities for a hybrid system. During the concurrent training of the SOMs, Oja links (associative Oja network) are created between the simultaneously most active nodes on the two maps. The outputs of the SOMs act as an input to the Oja network; therefore, the network not only learns to associate the two map regions but also learns the strength of the association—the higher the activation of the nodes in the two SOMs, the higher the strength. This hybrid multi-net system can operate as an image annotation system that, once trained, can annotate unannotated images (labeling). The system can be used equally as a keyword illustration system for image retrieval. In contrast to existing systems, particularly user relevance feedback systems, our approach only requires an exemplar set of images with keywords for the training process. The proposed framework is an extension of the technique presented in [8] with an application in gastrointestinal data. So, a comparative analysis has also been performed between these techniques, along with different combinations of various image and text features.

The proposed research has several benefits in the medical field. Medical judgment is subjective and may vary from physician to physician, thus requiring comprehensive expert annotations to reach a concord. Computer-aided diagnosis considerably improves the prognosis by assisting physicians in clinical diagnosis and treatment [9]. For instance, doctors can review similar old medical cases by retrieving the related images or linguistic cues for better decision making on new cases [10]. Moreover, the introduced system can be valuable to medical undergraduates and interns for studying and learning about a disease by relating images with technical phraseology [11]. Patients can perform an automatic diagnosis by themselves from the laboratory reports and accompanying pictures (for example, X-rays, CT and MRI scans) using a website or a mobile application [12]. Researchers can enhance this work by encompassing more modalities to design a robotic system to utilize in the medical area and assist mankind [13].

The remaining article is organized as follows: Sect. 2 discusses related works, introduction to traditional SOM, image, and text features, Sect. 3 explains the Hebbian and Oja learning, Sect. 4 describes the proposed endoscopy cross-modal retrieval method, experimental analysis and results are covered in Sects. 5, and 6 concludes the work.

Contributions

The significant contributions of the article are as follows:

1. The proposed hybrid self-organizing map (HSOM) technique associates the information from endoscopy images and collateral text to create an image annotation and retrieval system.

2. Oja learning rule has been utilized to integrate two self-organizing maps trained separately on image and text modalities.
3. The experimentation has been performed on the real gastrointestinal data obtained from a known gastroenterologist.
4. Comparative analysis has been performed between HSOM with Hebb rule and HSOM with Oja rule image–text integration networks.

2 Background

This section covers the recent works related to the proposed approach and also gives an introduction to the traditional techniques which have been used in the existing literature.

2.1 Related work

In neural computing literature, there are reports of multi-net systems that comprise neural networks where each becomes an expert on solving a task or part of a task, and their outputs are joined to form a unique output. Multi-nets should, in principle, deal with complex nonlinear problems that single-nets cannot solve [14]. Images in large and broadly categorized online collections or collections developed by specialist communities (medical images and scene of crime images [15]) are often accompanied by one or more keywords per image. These collections can, in principle, be indexed and queried using the visual features of an image using so-called content-based image retrieval systems.

Annotating images by an expert is costly, so the authors have proposed a technique based on a shallow network and deep learning for the automatic generation of semantic description for an image [16, 17]. This has led researchers to combine visual features and keyword indexing. However, the keyword annotation of images is carried out manually [18, 19]. For instance, multi-net systems have been used in forensic science with some success [15]. The linguistic descriptions, especially keywords, are created to capture the essence of meaning related to a given object or event. Prototype systems have been developed for annotating images with keywords based on probabilistic associations between visual features and the collateral keywords [20]. The development of the PICSOM system [21] and its various extensions [22] shows how, given a set of images labeled with keywords, a system of self-organized classifiers can be trained to detect the presence of keywords. The output is then used to label a test set using a sophisticated image similarity assessment. PICSOM draws upon the experience of the large document retrieval system WEBSOM and continues with the tradition of building modular SOM systems [23].

In [24], authors have studied image retrieval using visual words and their weighted average of triangular histograms (WATH). The bag-of-visual words (BoVW) enables to ignore the spatial features, and an order-less histogram of visual words can represent an image. The proposed method has reduced the semantic gaps for image features and image semantics and overfitting issues for more extensive dictionaries. The pre-trained VGG16 network has been utilized for image feature extraction in [25]. The authors utilized the VGG16 features along with multi-class support vector machine (SVM) to classify diseases in Eggplant. VGG16 feature extraction has also been utilized in classifying malicious software with the help of SVM classifier [26].

Deep transfer learning (VGG16, VGG19, and DenseNet1 21) and data augmentation techniques have been utilized in [27] to design a fingerprint pattern classifier capable of classifying six different categories. In [28], authors compare multiple deep learning pre-trained models such as VGG19, VGG16, Alexnet, Inceptionv3, and Resnet50 on the Gujarati food image classification task. Authors in [29] have utilized VGG16, along with six other pre-trained deep learning models for cotton weed recognition. VGG16 has been used as an image feature extractor in [30] for multispectral pedestrian detection. A deep supervised hashing approach based on selective pool feature map has been proposed in [31] for image retrieval and two pre-trained models, VGG16 and AlexNet, have been exploited. Zernike moments have been used for edge detection in the concrete surface roughness measurement method [32].

Paek et al. [33] integrated the images and texts by applying the statistical method *tf*idf* for text and *of*iif* for images and combining the two measurements. *tf*idf* is the term frequency multiplied by the inverse document frequency, a well-known approach for classifying text. Paek et al. created integrated feature vectors that are used for categorizing indoor/outdoor images and texts with an accuracy of 86%. In [34], authors have focused on understanding the semantic concept rather than traditional syntactic for clustering text records. TFIDF and K-means algorithms are used for document clustering. A multimodal semantics enhanced joint embedding technique has been proposed in [35]. Here, TFIDF features are integrated with LSTM for effective recipe-image cross-modal retrieval process. In [36], a PAN-LDA approach has been proposed where latent Dirichlet allocation (LDA) features have been utilized for analyzing the coronavirus disease data along with online news articles, generating a new set of features.

Latent semantic indexing (LSI), a recognized text document indexing method, has been adapted to work in conjunction with image features by other authors with varying degrees of success [37]. It has been argued that LSI cannot distinguish between different modalities in a joint feature space and tends to perform quite poorly compared to techniques that differentiate between different modalities, including cross-modal factor analysis [38].

Neuron-based spiking transmission and reasoning network (NSTRN) is proposed in [39], which is motivated by the neuron spike signals in the brain. Spiking activation function inspired feature sender encodes only vital information in images and text into binary codes, which lessens the transmission cost. The feature receiver has a recurrent design that learns long-term information by applying global and temporal attention blocks. To learn the joint image–text representations, a novel contrastive cross-modal knowledge sharing pre-training (COOKIE) approach has been introduced in [40]. A module in the proposed method comprises a weight-sharing transformer placed on the head of image and text encoders, which semantically aligns the visual and textual information. Contrastive learning has been incorporated to share knowledge between diverse models. Authors in [41] have introduced a unified framework with ranking learning (URL) for effective cross-modal retrieval tasks. The framework comprises a visual network, textual network, and interaction network. Visual and textual networks project the image and text features into their respective hidden spaces, and the interaction network impels the target image–text representation to associate in the shared space.

2.2 Introduction to traditional SOM

The self-organizing map is also popularly known as the *Kohonen map* after the name of its founder *Teuvo Kohonen* [42]. SOM is an unsupervised machine learning technique that maps the multi-dimensional data to (usually) a two-dimensional grid of neurons or nodes referred to as a map. Similar input instances are associated with nodes closer in the grid; however, the less similar ones are linked with farther nodes [43]. The fundamental idea behind SOM is that each input vector is associated with a SOM node that best matches it or the node that wins it (referred to as the best matching unit or BMU). The BMU's spatial neighbors in the map are also dragged toward that input vector, updating the shape of the map. A SOM node can act as the BMU for multiple inputs. SOM helps visualize the high-dimensional data by mapping it into a 2-D map and clusters similar data together. A traditional self-organizing map consists of two layers: (1) the input layer (which comprises the input instances); and (2) the output layer (which constitutes a grid of SOM neurons/nodes for mapping inputs). The step-by-step procedure of SOM learning is described in [44].

2.3 Image feature extraction

Endoscopy image feature extraction has been performed using Zernike moments (ZM) and a pre-trained VGG16 convolution neural network. The features have been chosen to represent an image's miscellaneous and significant properties. Zernike moments extract the prominent features of an image and proved their effectiveness in multiple applications recently [45–47]. A classification experiment has been performed with several pre-trained deep convolution neural networks for selecting the appropriate network for deep visual feature extraction. After feature extraction, images are classified into respective classes using a support vector machine (SVM) classifier. Out of all the deep model features, SVM gave the highest accuracy using VGG16 features, so it has been utilized in the proposed approach for image feature extraction. Table 1 shows the models chosen for experimentation (in MATLAB R2019a) along with the classification accuracy. VGG16 and the corresponding accuracy value are represented as bold in the table because of the highest value. VGG16 convolution neural network has been used for classification or visual feature extraction in many applications recently and has also been found to be effective [25, 26, 48].

Table 1 Different deep learning models chosen for experimentation (in MATLAB) along with classification accuracy (using SVM) to select the best accuracy model for image feature extraction

Sr.	Model	Accuracy
1	alexnet	0.7667
2	**vgg16**	**0.8333**
3	googlenet	0.6667
4	squeezenet	0.6667
5	inceptionv3	0.6333
6	densenet201	0.8
7	mobilenetv2	0.6667
8	resnet18	0.6333
9	resnet50	0.7667
10	resnet101	0.7
11	inceptionresnetv2	0.6

2.3.1 Zernike moments (ZM)

Zernike moments are continuous orthogonal moments that are least redundant, noise resilient, scale, rotation, and translation invariant [49]. They are derived from the complex Zernike polynomials proposed by *Frits Zernike* (optical physicist) [50]. These polynomials are defined within a unit disk over polar coordinate space. ZM are defined as the projections of an image function along real and imaginary axes which are convolved by an orthogonal function. Therefore, images are represented in diverse frequency components such as orders (along radial direction) and repetitions (along angular direction). For ZM calculation, an image is mapped onto a unit circle to transform the image center into the center of the circle using the outer circle mapping technique [51].

The process followed for ZM calculation along with image pre-processing is same as [8].

If an image function is represented by $f(r,\phi)$, the 2-D ZM with order p and degree of repetition q are calculated in Polar coordinate system as below [52]:

$$Z_{pq} = \frac{p+1}{\pi} \int_0^{2\pi} \int_0^1 f(r,\phi) V_{pq}^*(r,\phi) r dr d\phi \tag{1}$$

where $V_{pq}^*(r,\phi)$ is the complex conjugate of Zernike polynomials $(V_{pq}(r,\phi))$ which are evaluated as:

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

which satisfies $p \geq 0$, $0 \leq\mid q \mid\leq p$, $p- \mid q \mid$ = even, and $i = \sqrt{-1}$. r is the length of the vector from origin to the point (x, y) and θ depicts the angle between x-axis and the vector. Radial polynomials can be calculated as below:

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

2.3.2 VGG16

VGG16 network is developed by the *Visual Geometry Group* of the University of Oxford, and it is the winner of the 2014 ILSVRC object identification algorithm [53]. It is a deep convolutional neural network pre-trained on the ImageNet dataset comprising more than a million images. It is capable of categorizing the images into 1000 object classes. Hence, the network has learned rich feature representations for a variety of images. VGG16 takes the fixed input of 224 × 224 RGB image. Its network architecture consists of a total of 41 stacked layers. There are 16 layers with learnable weights: 13 convolutional layers and 3 fully connected layers. A kernel of 3 × 3 dimension is utilized for the convolution operation along with W and b (as learnable attributes), which are passed over the pixels x of an image, and it gives y as the output. The following equation simply represents the convolution task by the function:

$$y = f(Wx + b) \tag{4}$$

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

2.4 Text feature extraction

LDA and TFIDF features have been used for text vector creation in the proposed framework. The motive of text feature extraction is to choose the significant words in each text document that can uniquely identify it and represent the document with a vector for mathematical implementation. The feature extraction methods are described in the following sub-sections.

2.4.1 Latent Dirichlet allocation (LDA)

LDA is one of the popular and prominent techniques of topic modeling. The procedure followed for LDA feature extraction is same as given in [8]. LDA is represented as 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 [54]. A *word* can be represented as a fundamental element of discrete data and a vocabulary unit. A series of R words depicted as $\mathbf{w} = (w_1, w_2, \ldots, w_R)$ is referred to as a *document*, where w_r is r^{th} word in the sequence. A *corpus* is defined as a bunch of Q documents depicted by $C = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_Q)$. The three-level probabilistic graphical model of LDA is shown in Fig. 3. The inner plate in the model represents the recurrent topic and word choice in a document, whereas the outer plate signifies documents. The parameters γ and δ are corpus level parameters that are sampled one time while corpus generation. The symbol η depicts the variables at the document level that are sampled once in a document. The symbols z and w represent variables at the word level that are sampled once in a document for a single word.

Each document $\mathbf{w}$ in a corpus C follows a generative procedure in LDA, given below [54]:

1. Select $R \sim$ Poisson(ξ).
2. Select $\eta \sim$ Dir(γ).
3. For every word w_r in a document, select:

    (a) a topic $z_r \sim$ Multinomial(η).

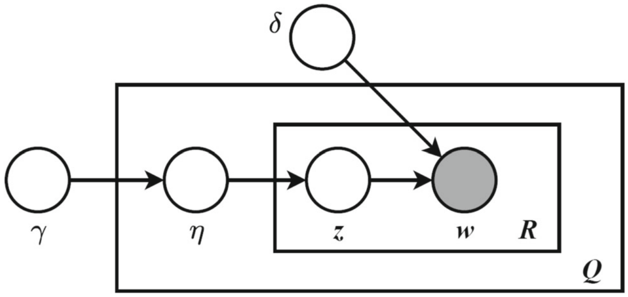


Fig. 3 Graphical model representation of LDA [54]

(b) a word w_r from $p(w_r|z_r, \delta)$, a multinomial probability conditioned on z_r topic.

A few assumptions that are formed in the basic LDA model are: (1) dimensionality of Dirichlet distribution is known and stable; (2) R is not dependent on other data generating 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 taken as a steady quantity that is to be determined.

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 following probability density on this simplex:

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

where γ depicts a k-vector with $\gamma_i > 0$ and $\Gamma(x)$ is a 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 $\mathbf{w}$ is evaluated as:

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

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

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

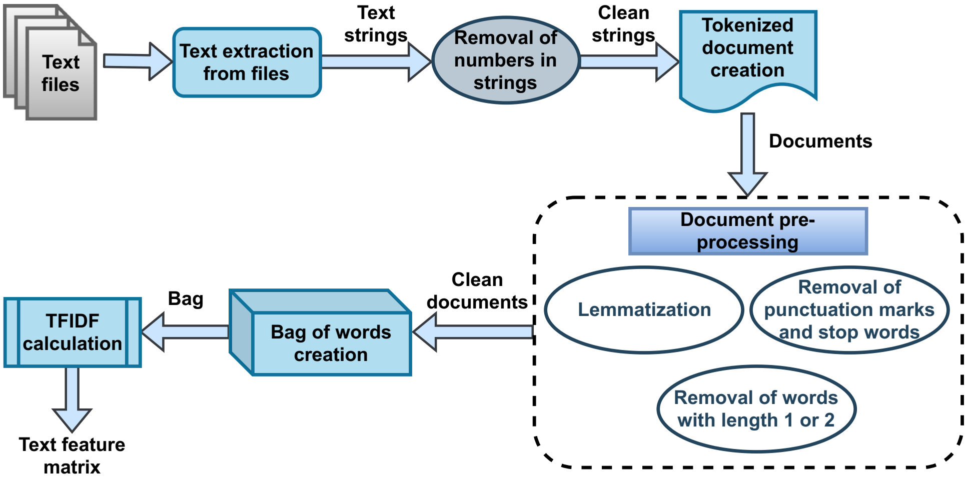
Afterward, the probability of a corpus can be determined by taking the product of the marginal probabilities of single documents:

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

2.4.2 TFIDF

Frequency-based metrics are typically utilized to construct vectors for sets of text documents. The vectors comprise information about the presence or absence of *significant* keywords. One of the widely used methods in document representation is called the *TFIDF* (term frequency–inverse document frequency) method that weighs the significance of a keyword, based on its overall frequency in a document set (term frequency) and the number of documents that have at least one instance of a given keyword. The two components of the TFIDF metric are computed over the entire corpus and significant keywords are then selected. High-value TFIDF terms basically indicate terms that appear to be significant for the specific document. The more frequent a term is in

Fig. 4 Process flow for TFIDF text feature extraction


Text files
Text extraction from files
Text strings
Removal of numbers in strings
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
TFIDF calculation
Text feature matrix

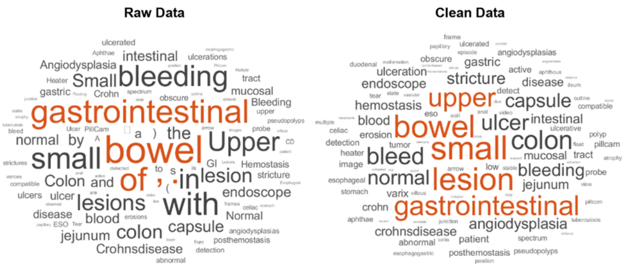


Fig. 5 Raw data versus cleaned data after pre-processing

a document and the less it appears in other documents the higher its weight. A term that appears in every other document has a zero weight. TFIDF weight ($tfidf_{i,j}$) of a token i in document j is given as [55]:

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{9}$$

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

Figure 4 represents the steps followed for TFIDF text feature extraction. The first step is to extract the text from the textual files and convert them into text strings, and then the numbers (if there are any) are removed from the strings. These strings are used to create tokenized documents which are then pre-processed by lemmatization and removing punctuation marks, stop words, and words with length 1 or 2. Afterward, a cleaned bag of words is created from the cleaned documents, and this bag is used for TFIDF feature extraction. Figure 5 shows the word clouds of raw and cleaned data (obtained after pre-processing).

3 Hebbian and Oja learning

The neuropsychologist *Donald Hebb* postulated regarding the learning of the biological neurons [56]:

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

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

$$\Delta w_{ij} = \eta \cdot a_i a_j \tag{10}$$

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

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

$$a_j = \sum_i w_{ij} a_i = \mathbf{a}^T \times \mathbf{w}_j \tag{11}$$

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

$$\Delta \mathbf{w}_j = \eta \mathbf{a} a_j = \eta \mathbf{a} \times \mathbf{a}^T \times \mathbf{w}_j = \eta Q \times \mathbf{w}_j \tag{12}$$

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

$$Q = \mathbb{E}_{\mathbf{a}}[\mathbf{a} \times \mathbf{a}^T] \tag{13}$$

Hence, Hebbian plasticity is learning strong weights to frequently co-occurring input elements. The simple Hebbian learning rule suffers from a severe issue. There is nothing to stop the connections from growing all the time, eventually leading to huge values. So, weights will keep on growing in size with time. Another term is required to balance this growth. A term depicting "forgetting" has been utilized in several neuron models where the weight value itself should be subtracted from the right-hand side. *Erkki Oja*, a *Finnish computer scientist* proposed a learning rule, known as *Oja rule*, which is a mathematical formalization of the Hebbian rule, such that a neuron learns to compute a principal component of its input stream over time [57]. The main idea behind this rule is to make the forgetting term proportional to the value of weight along with the square of the activity of the postsynaptic neuron. Oja rule normalizes the length of a weight vector by a local operation:

$$\Delta w_{ij} = \eta a_i a_j - \eta a_j^2 w_{ij} \tag{14}$$

$\eta a_j^2 w_{ij}$ is a regularization term. When the postsynaptic activity a_j or weight w_{ij} is too large, then the term cancels the Hebbian $a_i a_j$ part and decreases the weight. The new (next) weight can be calculated as the old weight ($w_{ij}(n-1)$) plus the change in weight (Δw_{ij}) as given in Eq. 15:

$$w_{ij}(n) = w_{ij}(n-1) + \Delta w_{ij} \tag{15}$$

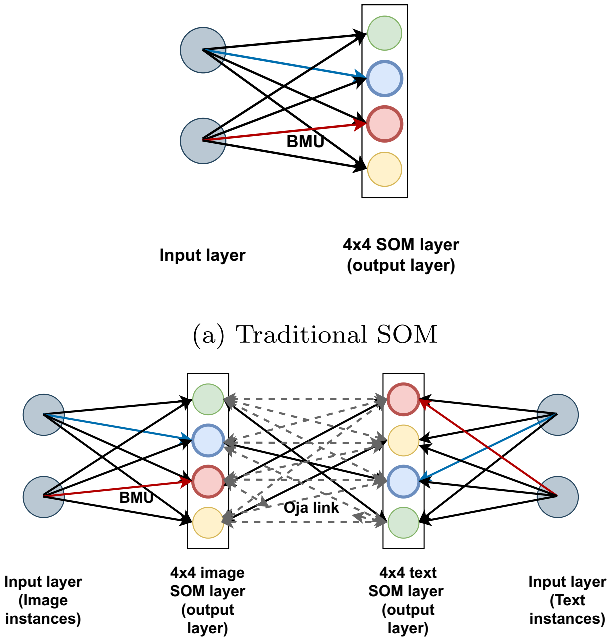
4 Oja learning-based cross-modal retrieval

4.1 Problem formulation

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

4.2 Proposed HSOM technique exploiting Oja rule

The proposed approach aims to associate two separately trained traditional self-organizing maps (on diverse modalities) using improved Hebb links or Oja links, creating a hybrid SOM model which can be utilized for endoscopy image annotation and retrieval. Figure 6 demonstrates the difference between the traditional SOM (Fig. 6a) and hybrid SOM (HSOM) (Fig. 6b) using two data instances in the input layer. HSOM associates two separately trained traditional SOMs on different modalities using Oja links, as shown in the figure. In the proposed study, one SOM is dedicated to the image modality (dubbed image SOM) and the other to the text modality (dubbed text SOM). These two SOMs are integrated using a third network known as the improved Hebbian network or Oja network that connects each neuron (similar cluster of images) in image SOM with every neuron (similar cluster of texts) in the text SOM. Oja network is inspired by the Oja learning rule described in Sect. 3. Neurons in the trained SOMs that are synchronously highly active while training are associated via the Oja network. This network has been used for the association to enhance the connections between the SOMs when respective nodes in them activate in response to an input text or an image. If the size of the text SOM is $m \times n$ and image SOM is $p \times q$, then the size of the connecting network would be $m \times n \times p \times q$. In the proposed study, the size of both the image and text SOM is 4×4, so the size of the Oja network is 16×16.



(a) Traditional SOM

(b) Hybrid SOM

Fig. 6 Demonstration of **a** traditional SOM; and **b** hybrid SOM where two separately trained traditional SOMs are integrated using Oja links

For implementing the proposed approach, image and text features are retrieved as described in Sects. (2.3, 2.4). Two independent SOMs net_T and net_I of dimension 4×4 are trained for texts and images correspondingly, and the SOM neuron numbers are also obtained corresponding to each instance depicted as $classes_T$ and $classes_I$ matrices. The SOM node weights represented by $nodeWeights_I$ and $nodeWeights_T$ are obtained for both image and text SOM for further experimentation. Let $winnersMatrix_T$ and $winnersMatrix_I$ be the weight vector of the winner node of each text and image input instance, respectively. Afterward, Euclidean distance has been calculated between each input vector and the corresponding winner node weight vector, and the results are depicted as one-dimensional matrices $woja_I$ and $woja_T$ individually. Now the training of the Oja network is performed as per Eq. 16, and the Oja link weights (represented by $ojaLink$ matrix) keep on updating for each input instance for the whole training process. All the nodes of net_I are connected with all the nodes of net_T in the Oja network. However, the strength of the Oja bond is determined by the Oja link weight.

$$\begin{aligned} & ojaLink(classes_I(i), classes_T(i)) \\ & \quad = ojaLink(classes_I(i), classes_T(i)) \\ & \qquad + (\alpha * woja_I(i) * woja_T(i) \end{aligned}$$

$$-\alpha * woja_T(i) * ojaLink(classes_I(i), classes_T(i)) * woja_T(i)); \qquad (16)$$

where $1 \leq i \leq length(classes_I)$ and the α represents the learning rate whose value has been chosen to be 0.001 after experimentation on values given in the set $\{0.1, 0.01, 0.001\}$. After the network training, two vectors $Anet_I$ and $Anet_T$ of size 16 are created such that $Anet_I$ will have the node numbers of net_T having the highest Oja link weight where each index of the $Anet_I$ vector represents the node number in net_I. Similarly, $Anet_T$ comprises the node numbers of net_I SOM. For model testing using an image input, firstly, the input image is clustered in a suitable node in the image SOM as per the similarity. Afterward, the corresponding linked node in the text SOM is found, and the clustered instances from both the image and text nodes are retrieved. Algorithms (1 and 2) display all the steps performed for implementing the proposed method. Figure 7 shows the pipeline diagram of the proposed system for cross-modal gastrointestinal images and text retrieval.

5 Experimental analysis

This section includes all the implementation details of the proposed framework and the results obtained. The implementation has been performed in *MATLAB R2019a* on Windows 10 system with 8GB RAM.

5.1 Dataset

A real endoscopy dataset has been composed for this study by visiting a known gastroenterologist. He provided the related keywords corresponding to each of the images, which can be utilized as the other modality for training. Total 300 gastrointestinal data instances (image–text pair) have been collected. The dataset has a ratio of 180:120 for healthy and sick cases. Figure 8 illustrates the word clouds of normal and sick categories and Fig. 9 shows an image–text pair instance from the dataset. All the images are of size 256 × 256, and the data incorporates four diverse categories of in vivo gastral images: Upper (normal and bleeding) and Lower (normal and bleeding) as shown in Fig. 10. The upper gastrointestinal tract includes *esophagus and stomach*, lower includes *small bowel and colon*. These areas have been further divided into normal (180) or bleeding (120) instances. In upper GI tract, normal = 42, bleeding = 54; in lower GI tract, normal = 138 and bleeding = 66 instances. Table 2 provides information regarding the endoscopic images in the dataset. The plot in Fig. 11 shows the red color intensity comparison of healthy and sick images. It can be visualized that sick images contain more redness than healthy images.

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

```
INPUT: $E_{train}$ and $E_{test}$
OUTPUT: Trained $net_I$ and $net_T$ SOMs, retrieval of matched images and text corresponding to text and images in $E_{test}$, respectively
procedure IMAGE FEATURE EXTRACTION
    Input all images
    Resize the images to 224 × 224 × 3 ▷ Defined image input size for VGG16 input
    Extract the deep features of 1000 dimension from fc8 layer of VGG16 network
end procedure
procedure TEXT FEATURE EXTRACTION
    Input all text files
    Removal of numbers from each text
    $cleanedDocuments \leftarrow tokenizedDocument(text)$ ▷ Create tokenized documents from the text
    Perform lemmatization
    Remove punctuation marks, stop words and words with $length \leq 2$
    $cleanedBag \leftarrow bagOfWords(cleanedDocuments)$ ▷ Create a bag-of-words from cleaned documents
    Calculate TFIDF score from the cleaned bag using Eq. 9
end procedure
procedure HSOM BASED CROSS- MODAL RETRIEVAL USING ENDOSCOPY DATA
    Load $E_{train}$ and $E_{test}$
    $dimension1 \leftarrow 4, dimension2 \leftarrow 4$ ▷ Dimensions of both image and text SOM
    $net_I \leftarrow selforgmap([dimension1 dimension2])$ ▷ Configure image SOM with default parameters except dimensions
    $net_T \leftarrow selforgmap([dimension1 dimension2])$ ▷ Configure text SOM with default parameters except dimensions
    $net_I \leftarrow train(net_I, I_{train}), net_T \leftarrow train(net_T, T_{train})$ ▷ Training of maps
    $classes_I \leftarrow vec2ind(net_I(I_{train})), classes_T \leftarrow vec2ind(net_T(T_{train}))$ ▷ Retrieving node number for each input instance
    for $i \leftarrow 1$ to $length(classes_I)$ do ▷ Winner node weight matrix corresponding to image input instances
        $winner_I \leftarrow classes_I(i)$
        $winnersMatrix_I(:, i) \leftarrow nodeWeights_I(winner_I, :)'$
    end for
    for $i \leftarrow 1$ to $length(classes_T)$ do ▷ Winner node weight matrix corresponding to text input instances
        $winner_T \leftarrow classes_T(i)$
        $winnersMatrix_T(:, i) \leftarrow nodeWeights_T(winner_T, :)'$
    end for
    for $i \leftarrow 1$ to $length(classes_I)$ do ▷ Euclindean distance calculation
        for $j \leftarrow 1$ to $imageVectorDimension$ do
            $woja_I(i) \leftarrow woja_I(i) + (winnersMatrix_I(j, i) - input_I(j, i))^2$
        end for
        for $j \leftarrow 1$ to $textVectorDimension$ do
            $woja_T(i) \leftarrow woja_T(i) + (winnersMatrix_T(j, i) - input_T(j, i))^2$
        end for
        $woja_I(i) \leftarrow sqrt(woja_I(i))$
        $woja_T(i) \leftarrow sqrt(woja_T(i))$
    end for
    for $i \leftarrow 1$ to $length(classes_I)$ do
        Train the improved Hebbian (Oja) network using Eq. 16
    end for
    Follow Algorithm 2 for creation of $Anet_I$ and $Anet_T$
    Cluster $I_k \in net_I$ and $T_k \in net_T$ where $(I_k, T_k) \in E_{test}$ and $k \in [1, N_2]$
    Refer $Anet_I$ and $Anet_T$ to find the corresponding Oja link node
    Retrieve results from the found node
end procedure
```

5.2 Evaluation metrics

The two evaluation metrics utilized in this study for performance analysis are as follows:

1. *Mean average precision (MAP)*: It is a prevalent metric for evaluating the performance of a cross-modal information retrieval system. This metric checks whether the

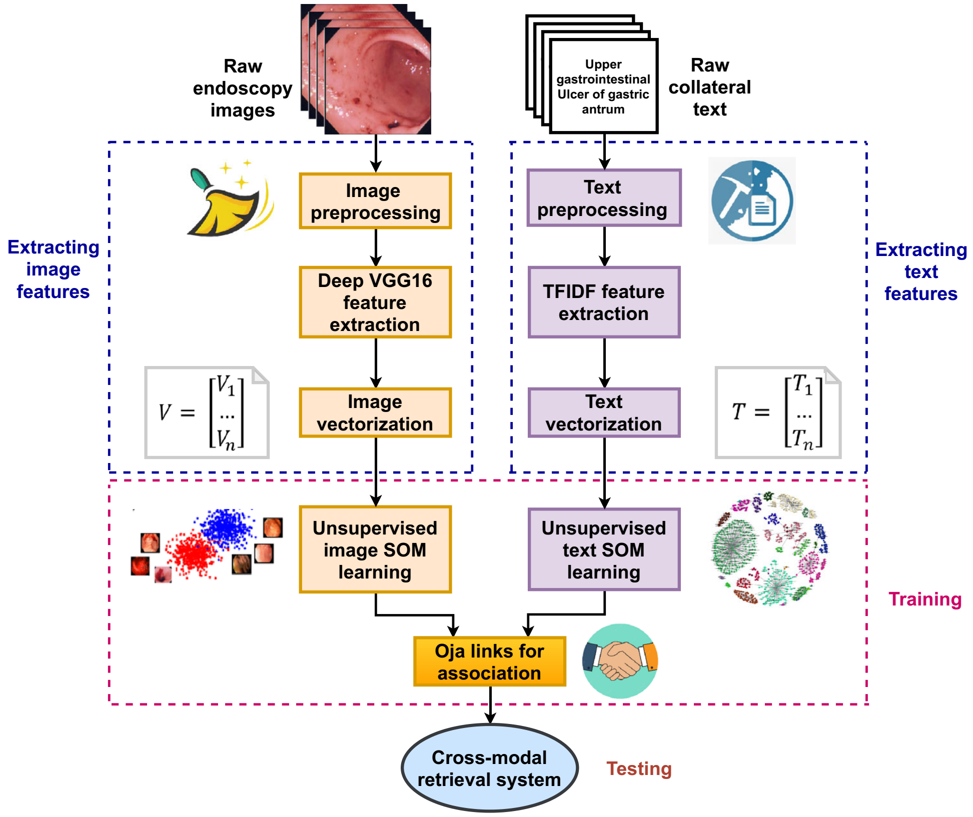

Raw endoscopy images
Upper gastrointestinal Ulcer of gastric antrum
Raw collateral text
Extracting image features
Image preprocessing
Deep VGG16 feature extraction
$V = \begin{bmatrix} V_1 \\ \cdots \\ V_n \end{bmatrix}$
Image vectorization
Text preprocessing
TFIDF feature extraction
$T = \begin{bmatrix} T_1 \\ \cdots \\ T_n \end{bmatrix}$
Text vectorization
Extracting text features
Unsupervised image SOM learning
Unsupervised text SOM learning
Training
Oja links for association
Cross-modal retrieval system
Testing

Fig. 7 Process flow of image and text training for the proposed hybrid cross-modal retrieval system

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

```
INPUT: Trained Oja 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 Oja 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 $ojaLink(i, j) > maxtemp$ then   ▷ Checking for the maximum ojaLink weight
                $maxtemp = ojaLink(i, j)$
                $maxindex = j$
            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 $ojaLink(j, i) > maxtemp$ then   ▷ Checking for the maximum ojaLink weight
                $maxtemp = ojaLink(j, i)$
                $maxindex = j$
            end if
        end for
        $Anet_T(i) = maxindex$
    end for
end procedure
```

retrieved outcome is of the same class as query (relevant) or not (irrelevant) [58]. MAP is defined as the mean of the measured average precision (AP) over all the test queries.

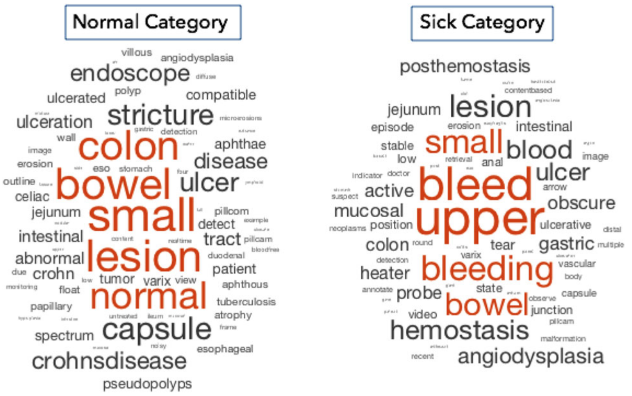


Fig. 8 Word clouds of normal (healthy) and sick classes

Given an input test query (a text or an image) and a set of corresponding retrieved outcomes Y, AP can be evaluated as:

$$AP = \frac{1}{R} \sum_{y=1}^{Y} P(y) rel(y) \tag{17}$$

where R depicts the ground truth positives or the number of relevant outcomes in the retrieved outcomes [59], $P(y)$ represents the precision of top y retrieved results, and Y has a different value for each test (image/text) instance

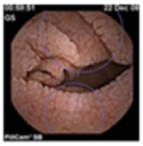
Image	Collateral text
	PillCam Small Bowel gluten refractory celiac disease

Fig. 9 Sample dataset instance: a gastral image and collateral text

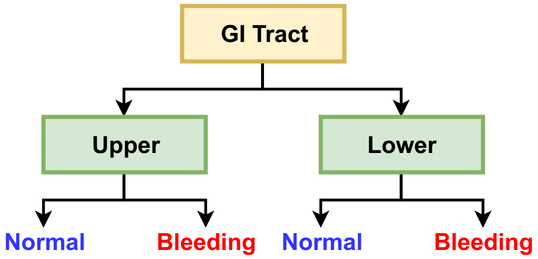


Fig. 10 Dataset of 300 images with collateral text has been divided into upper (esophagus, gastric) and lower (small bowel, colon) sections. Upper: normal = 42, bleeding = 54; Lower: normal = 138 and bleeding = 66

Table 2 Description of images in dataset

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

as the total number of obtained results are different for each test query. If the y^{th} retrieved result is relevant then $rel(y) = 1$ and otherwise 0. MAP score is defined as follows:

$$\mathrm{MAP} = \frac{1}{N}\sum_{n=1}^{N} AP \qquad (18)$$

where N signifies the number of queries. The higher the MAP score, the better the algorithm.

2. *Accuracy*: Accuracy is the common metric used for evaluating any algorithm. We define the accuracy of retrieval as the number of correctly retrieved items of a category, divided by the total number of items in the category.

5.3 Single SOM illustration

We begin by looking at the performance of a single SOM to deal with the categorization of images. Figure 12 is a collection of 10 healthy and 10 sick gastral images along with their captions. A simple SOM has been executed for individual image and text clustering for the 4×4 map as shown in Figs. 13 and 14, respectively. The image category has subtle visual features that are shared among each member: color, texture, shape, and edges. Hence, the clustering of gastral images together is no accident as depicted in Fig. 13.

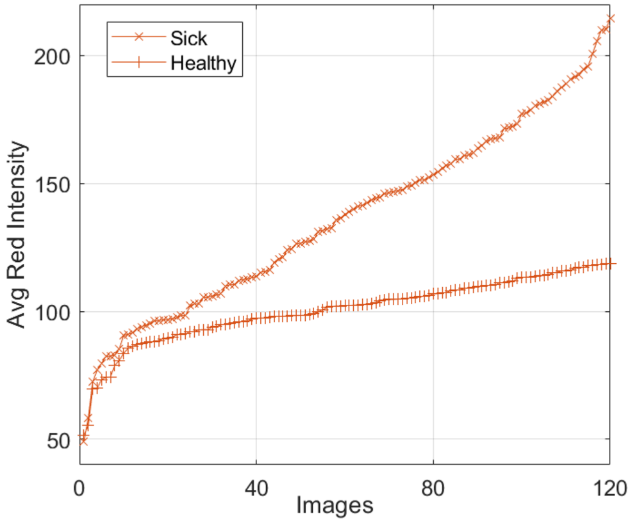


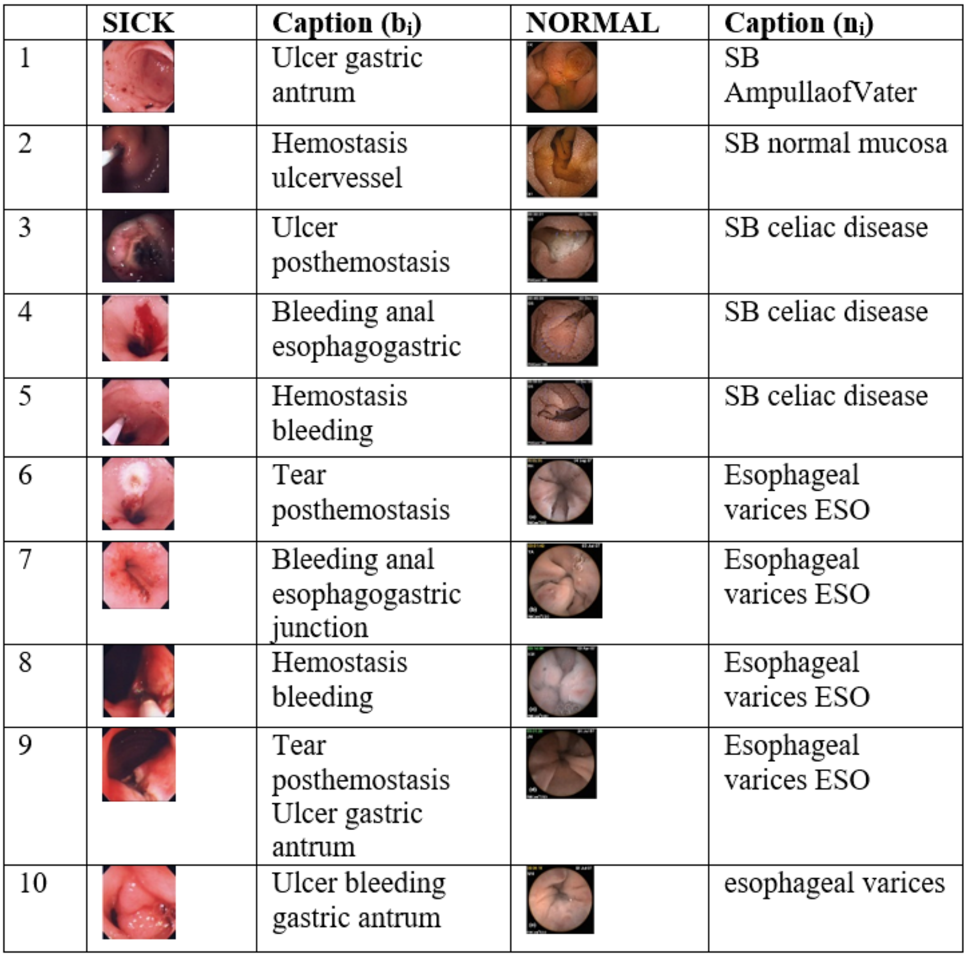
Fig. 11 Sorted average red pixel intensities for healthy and sick images. The sick images comprise more redness than healthy images

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

For an image-to-image retrieval, the overall recognition accuracy for single SOM has been found to be 77% as shown in Table 3. The results shown are the average of 5 runs. The keyword categorizing SOM performs a relatively simple task in that it matches keywords with keywords, and the accuracy of the retrieval for 30 keyword vectors presented 5 times was almost 100%. It should be noted here that choice of images and the assignment of categories by dataset indexers varied considerably in detail and perhaps in accuracy, suggested by the variation in the number of keywords, and some labels described visual features like ulcer, Angioectasia, and so on.

5.4 Testing HSOM exploiting basic Hebb rule

The hybrid system is trained using ZM (image representation), LDA (text representation), and Hebbian network (integration of two SOMs) as explained in [8]. The trained system is capable of retrieving both the same as well as different modality than the query modality. However, the main focus of a cross-modal retrieval is to analyze the accuracy of a system based on the efficiency of retrieving different modality than the query modality. So, the two tasks that are being automated here are: (1) *Auto-annotation*, and (2) *Auto-retrieval*



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

Fig. 12 A small example of 20 images and 20 captions. Here, b_i and n_i means the corresponding captions of bleeding and normal examples

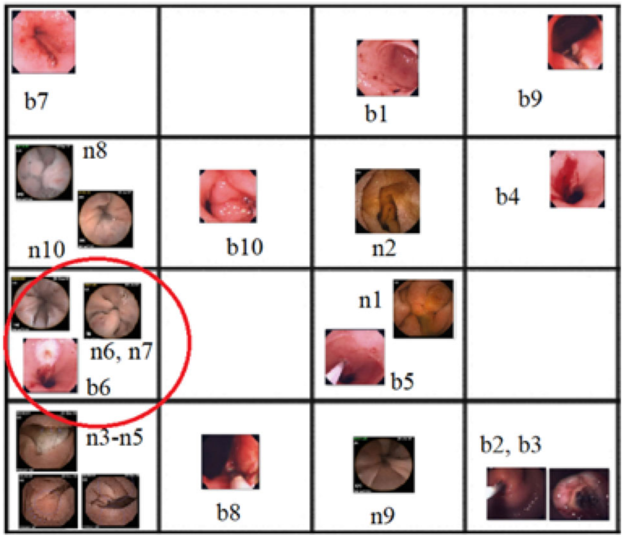
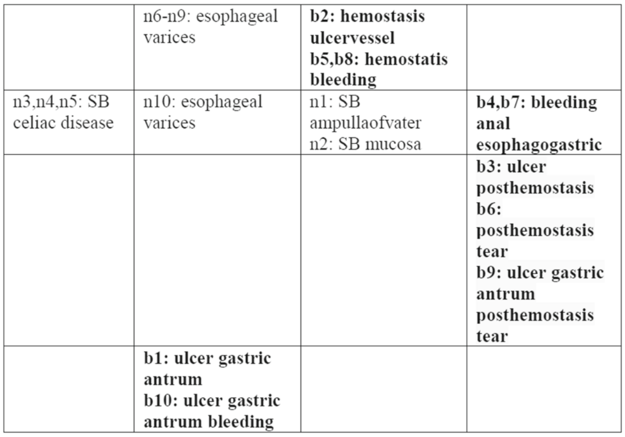


Fig. 13 Clusters created by the image features (ZM) in the SOM on 20 images (Fig. 12) for 4 × 4 grid of SOM. The circle shows overlapping node with normal and bleeding images



	n6-n9: esophageal varices	**b2: hemostasis ulcervessel** **b5,b8: hemostatis bleeding**	
n3,n4,n5: SB celiac disease	n10: esophageal varices	n1: SB ampullaofvater n2: SB mucosa	**b4,b7: bleeding anal esophagogastric**
			b3: ulcer posthemostasis **b6: posthemostasis tear** **b9: ulcer gastric antrum posthemostasis tear**
	b1: ulcer gastric antrum **b10: ulcer gastric antrum bleeding**		

Fig. 14 Clusters created by the text features in the 4 × 4 SOM for the sample data shown in Fig. 12. *n* and *b* means normal and bleeding instances, respectively. The keywords next to *n* and *b* are the keywords corresponding to that particular instance

Table 3 Accuracy scores of different retrieval operations using single SOM and hybrid SOM on endoscopy data

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

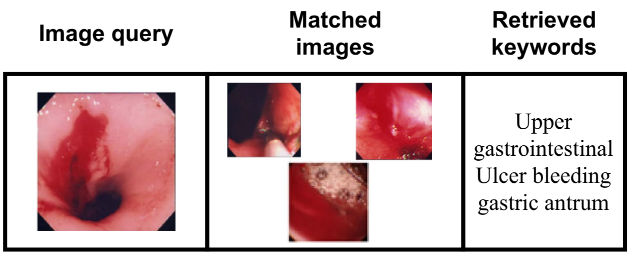


Fig. 15 Auto-annotation testing results using a single image query

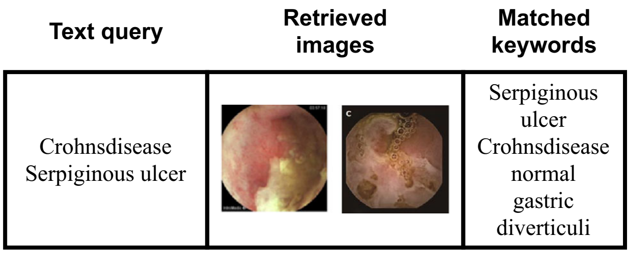

Text query
Retrieved images
Matched keywords
Crohnsdisease Serpiginous ulcer
Serpiginous ulcer Crohnsdisease normal gastric diverticuli

Fig. 16 Testing results for auto-retrieval on a test instance. Only 2 images have been shown to simplify the illustration

1. *Auto-annotation*: The trained hybrid system, comprising an image categorization system and a keyword categorization system connected via a Hebb link, is presented with an unseen image feature vector. The image SOM component is activated by the unseen image vector, and the system finds the best matching units (BMUs). The active nodes then activate nodes in the collateral keyword SOM via the Hebbian links. The hybrid system determines the BMUs. The auto-annotation process on a test instance is shown in Fig. 15. The trained system also retrieves the images matching with the query image. Only a few retrieved results are shown in the image for simplification. The average accuracy obtained for auto-annotation task is 83% (Table 3).
2. *Auto-retrieval*: In this case, the hybrid system is given an unseen query keyword vector. The system matches the keywords with the pre-stored nodes in a keyword categorizing SOM. The identification of the place in which query-related keywords are concentrated then allows the search for collateral images whose features have been 'learned' by a SOM. An example of the auto-retrieval process in Fig. 16. Given a text query, the matched images are also retrieved along with the keywords. The average accuracy is 85% for image retrieval task (Table 3).

Results reported for the hybrid SOM method in Table 4 represent the average values of class-wise accuracy over the test images for different retrieval tasks. The bold accuracy values depict the best value of accuracy on that particular data class and the retrieval task. The chart in Fig. 17 represents the category-wise accuracy scores for various retrieval tasks. It can be observed from the chart that the best accuracy is obtained in case of both *Upper GI—Bleeding* and *Lower GI—Bleeding* categories for *T2I* and *I2T* retrieval tasks, respectively.

5.5 Testing HSOM exploiting Oja rule

The step-by-step procedure of this approach (dubbed $HSOM\_OJA$) is mentioned in *Algorithm* 1. Here, the VGG16 deep features are extracted from the endoscopy images, and TFIDF features represent collateral text. Then, the separate SOMs are trained using these visual and textual features. The two SOMs are integrated using an Oja network or improved Hebbian network following the Oja learning rule to make a final cross-modal system that can be utilized for gastrointestinal image annotation and retrieval. The implementation details are given in the following sections. The final results obtained using this $HSOM\_OJA$ approach are compared with the results obtained using HSOM exploiting the Hebb rule, ZM, and LDA (dubbed $HSOM\_HEBB$).

5.5.1 Parameter settings

The parameter values of the different methods utilized in the implementation are chosen to enhance the system's overall performance. Order 5 has been chosen for the ZM features used in $HSOM\_HEBB$, generating a 12d image vector. Perplexity and time analysis has been performed to determine the appropriate value for the total topics created in the LDA model for textual representation. This analysis has been represented in Fig. 18 in the form of a plot. *Perplexity* is the statistical measure of how well a probability model predicted a sample. The number of topics needs to be selected such that the perplexity value is minimized. The lesser the perplexity value, the better the selection of the number of topics. However, the convergence time of the LDA model also needs to be considered along with perplexity because, with an increase in the number of topics, LDA may require more time to converge. So, to critically analyze this trade-off, both the perplexity score and the elapsed time are plotted simultaneously against the number of topics as depicted in Fig. 18. As per the figure, the best choice for the total number of topics is 14, which has been selected for the LDA model training, and it generates the 14d text vector corresponding to each

Table 4 Comparison of the accuracy for the 4 categories of images (I) and collateral text (T)

Category	I2I (%)	T2I (%)	I2T (%)	Avg (%)	Best task
Upper GI—Normal	77	80	**87**	81	I2T
Upper GI—Bleeding	73	**93**	80	82	T2I
Lower GI—Normal	77	77	**80**	78	I2T
Lower GI—Bleeding	80	83	**93**	85	I2T

Bold numbers indicate the best value of accuracy

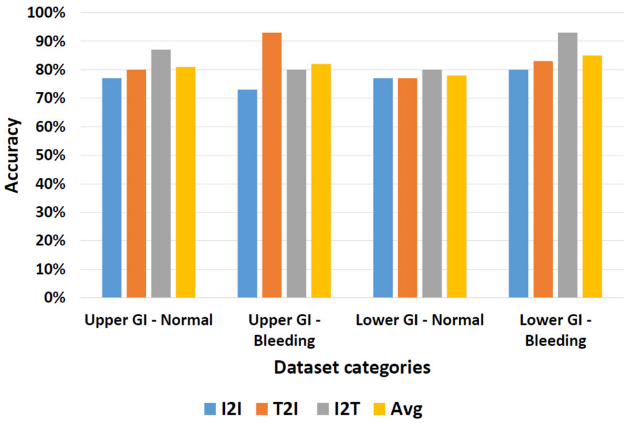


Fig. 17 Class-wise performance chart on different retrieval tasks based on accuracy scores

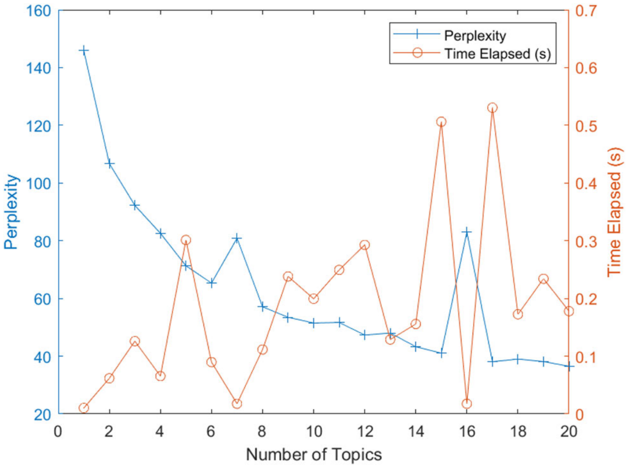


Fig. 18 Perplexity and time analysis to determine an appropriate number of topics for the LDA model in endoscopy data

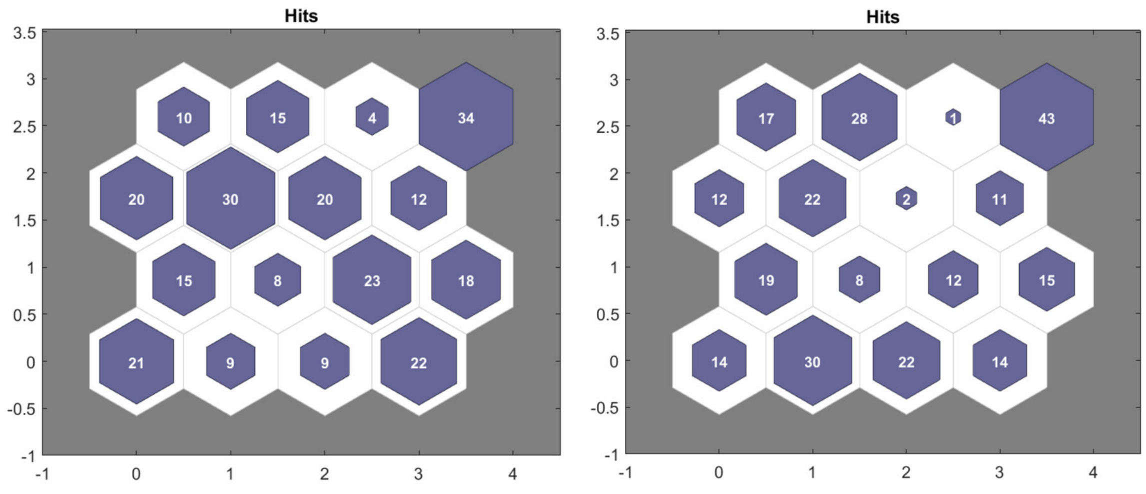
textual instance. For extraction of deep image features in the $HSOM\_OJA$ technique, a pre-trained VGG16 convolution neural network has been used with default parameters. The features have been extracted from the $fc8$ fully connected layer of the model generating a 1000d feature vector for image representation.

5.5.2 Model training

Figure 19 shows an example of the train data distribution after independent training of image and text SOM. The hexagon represents the neuron or a node in the SOM, and the number written inside each node depicts the total number of train instances clustered in that particular SOM node. These SOMs are trained using 1000d VGG16 visual features and 14d LDA (latent Dirichlet allocation) textual features. The corresponding SOM figures depicting neighbor distances are demonstrated in Fig. 20. Red lines connect SOM nodes (depicted as blue hexagons) to the neighbors. The darker the shade of the color in the red line section, the more is the distance between the adjacent nodes. The prominent tuning parameters chosen for image–text SOM training (after several experiments) in MATLAB are given in Table 5. The learning rate for the Oja network has been chosen by training multiple times at different values {0.1, 0.01, 0.001} and analyzing the average image–text query MAP score. After experimentation, 0.001 has been found to be an appropriate learning rate for the endoscopy dataset.

5.5.3 Results

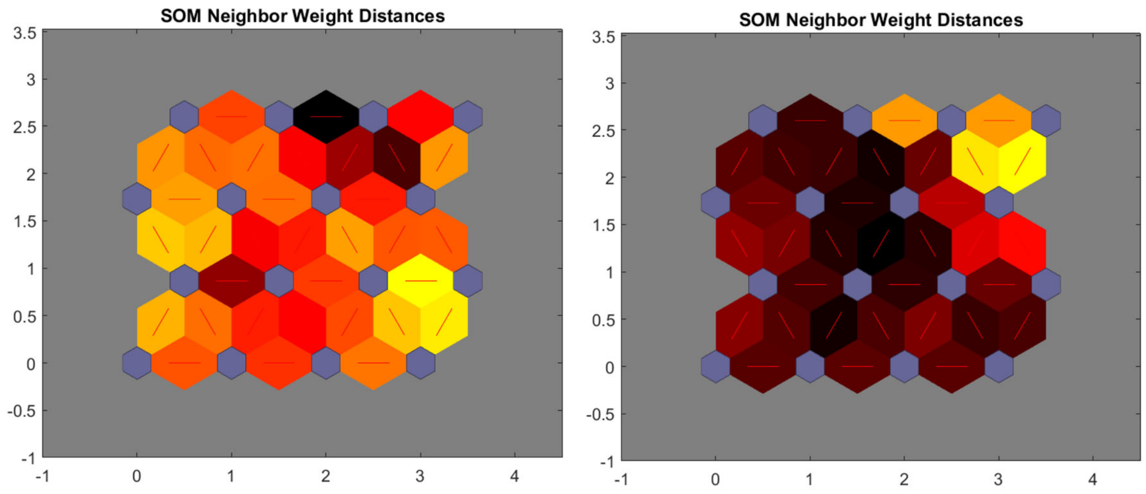
Table 6 shows the comparative analysis of different techniques based on the MAP score performance metric. Diverse combinations of image and text features have been utilized along with the two SOM association network weight updation rule. The table consists of five columns. *Rule* column represents the weight updation rule followed for training the image and text SOM integration network, such as the Hebbian and Oja learning rule. The columns *Image features* and *Text features* represent the various image and text representation methods used for the experiments, which are 12d ZM, 1000d VGG16, 14d LDA, and 323d TFIDF features. MAP_I2T and MAP_T2I columns represent the MAP score values for image-to-text retrieval (image annotation) and text-to-image retrieval (image retrieval), respectively. It can be observed that the Oja network training, along with VGG16 image features and LDA text features, is showing the best performance in image-to-text (I2T) retrieval operation; however, it is showing the worst performance for the T2I task. So, it can be deduced that the combination of VGG16 and LDA along with the Oja rule is not good for the overall



(a) Trained image SOM

(b) Trained text SOM

Fig. 19 Input train data distribution after individual SOM training for endoscopy data



(a) Trained image SOM

(b) Trained text SOM

Fig. 20 Neighbor distances among respective SOM nodes after training for endoscopy dataset. Darker shade denotes larger distance

Table 5 SOM parameters chosen for experimentation (in MATLAB) using endoscopy data

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

Table 6 MAP score comparison of different combinations of methods on endoscopy dataset

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

The best MAP scores in respective tasks are highlighted in bold

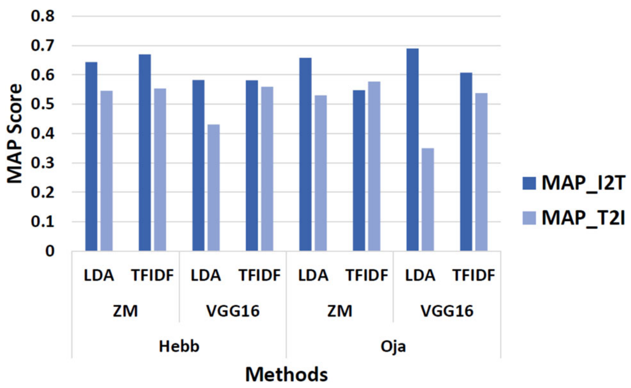


Fig. 21 Performance analysis of different methods based on MAP score

system performance. Oja rule with ZM image representation and TFIDF text representation is giving the best performance for text-to-image retrieval (T2I) task. The average (average of I2T and T2I MAP scores) system performance is best in the case of ZM image features, TFIDF text features, and Hebbian learning. This analysis aims to show the importance of choosing the methods wisely, as per the required operation and the application area. The performance (based upon the MAP score values) of various utilized methods can be visualized in the form of a bar chart in Fig. 21.

Table 7 displays the MAP score comparison of the proposed approaches with other state-of-the-art techniques. *Oja_ZM_TFIDF* depicts the application of the Oja rule with ZM image features and TFIDF text features, and similarly, *Oja_VGG16_LDA* represents the Oja rule with VGG16 image features and LDA text features. The best MAP scores are highlighted in bold. The sorted average precision results obtained using the Oja rule for all the test queries (image query for I2T and text query for T2I operation) are demonstrated in Fig. 22 in the form of curves. Four separate curves depict different combinations of image and text features utilized for the experimentation and SOMs associated with using the Oja network.

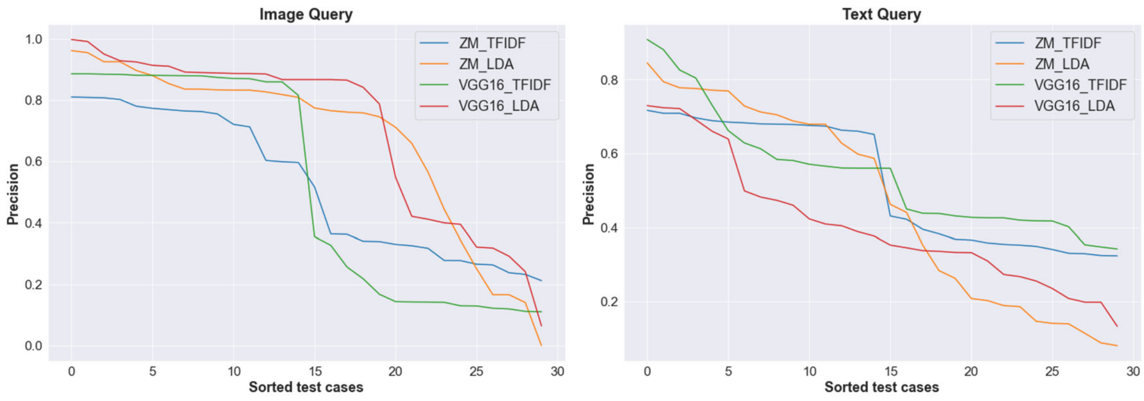
Table 7 MAP score comparison of the proposed approach with other state-of-the-art techniques

Method	MAP_I2T	MAP_T2I
URL [41]	0.5381	0.554
NSTRN [39]	0.6428	0.5519
COOKIE [40]	0.5245	0.562
Oja_ZM_TFIDF	0.5472	**0.5767**
Oja_VGG16_LDA	**0.6897**	0.3498

5.6 Discussion

The below reasons justify the effectiveness of the proposed approach:

1. Zernike moments extract the global image features and are robust shape descriptors. They are noise resilient, rotation, scaling, and translation invariant, and provide the least redundant features [60]. VGG16 features have been extracted from the pre-trained model, which has been chosen after comparison with several other pre-trained models.
2. LDA features divulge the inter and intra-document statistical structure. They provide well-defined inference procedures for even unseen documents [54]. Moreover, the number of topics in the LDA model has been chosen as per perplexity analysis.
3. SOM is a robust clustering algorithm capable of clustering massive datasets. It mimics the working of the human brain [42]. Moreover, SOM has lately been quite efficacious in a number applications such as energy sustainability [61], unveiling the comorbidities of chronic diseases [62], analysis of long-term evolution of groundwater hydrochemistry [63], intrusion detection of network viruses [64], and assessing public opinion [65]. The Hebb and Oja rules are also influenced by the biological systems [56].
4. The performance of a DL technique is greatly influenced by its design which incorporates training strategies, layer depth, and window size [66]. Moreover, it may be impractical to train a pre-trained model from scratch because it would require knowledge of several model parameters and layer modifications, which has a high computation cost [67].
5. Deep learning (DL)-based approaches usually require massive datasets (in millions) for model training [68]. The dataset in the proposed work contains fewer instances, which may result in overfitting.
6. Several deep convolutional neural networks (CNN) processes do not imitate the biological processes that occur inside the brain [69]. For instance, as per neuroscience, the brain does not have operations like backpropagation



(a) Image-to-text (b) Text-to-image

Fig. 22 Average precision results for all test queries obtained using Oja rule

and stochastic gradient descent optimization, which are the basis of deep CNN training. However, Hebb's learning rule better emulates the working of the human brain.

7. Selecting suitable feature extraction algorithms as per the modalities also plays a significant role in model effectiveness [3]. So the feature extractors have been chosen carefully in the proposed study and with appropriate parameter values to represent the modalities in the best promising way.

6 Conclusion

The proposed study introduced new ways of indexing and querying image collections by training neural computing systems with images accompanied by collateral texts. The characteristic visual features of the image collection are extracted using pre-trained VGG16 deep convolution neural network and Zernike moments. The characteristic linguistic features of the collateral texts are extracted using TFIDF and LDA, well-known text representation methods. The images and keywords were categorized synchronously but separately using an unsupervised clustering algorithm, Kohonen self-organizing feature maps. SOMs learn to categorize unseen images and collateral texts and concurrently, during uni-modal learning, an Oja link or Hebb link is established between the most active nodes in the two uni-modal maps: This is the basis of our claim that we use multimodal features to train neural networks and during the training to establish cross-modal connections between the two maps through another unsupervised network, the improved Hebbian learning network or Oja network. Multi-net architectures have been reported in the literature. Still, one seldom sees the establishment of cross-modal links during training and multi-sensory enhancement during testing (or retrieval) to compensate for the paucity of information in one mode. Our auto-annotation system's overall accuracy and MAP score are encouraging, although image semantics need to be considered more carefully in the auto-illustration system.

A few of the shortcomings of the proposed study are: (1) It has been validated on a small medical dataset; (2) it is challenging to obtain the medical dataset due to privacy issues and also of considerable size; and (3) the in vivo images suffer from moving cameras, image compression, optimal intensity of light, and occlusions; (4) the performance of the proposed cross-modal system is directly proportional to the details depicted by the respective captions of underlying images. Data augmentation techniques can be applied to refine the results in the future, and modified versions of ZM and LDA can be utilized. More associative learning approaches should be explored for integrating multiple modalities, and the technique can be applied in diverse application areas to show the reliability of the proposed framework.

Acknowledgements The authors are thankful to (a) Dr. Sunil Arya, Gastroenterologist at Leela Bhawan Patiala & Dr. G.S. Sidhu at Max Hospital Mohali, India, for the dataset and technical feedback; and (b) Professor Khurshid Ahmad, Trinity College Dublin, Ireland, for advice and research direction.

Data Availability Data sharing is not applicable to this article due to medical data privacy and the data owner may make it available based on individual request.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

1. Zhang, D., Islam, M.M., Lu, G.: A review on automatic image annotation techniques. Pattern Recogn. **45**(1), 346–362 (2012)
2. Dutta, A., Verma, Y., Jawahar, C.: Automatic image annotation: the quirks and what works. Multimed. Tools Appl. **77**(24), 31991–32011 (2018)
3. Kaur, P., Pannu, H.S., Malhi, A.K.: Comparative analysis on cross-modal information retrieval: a review. Comput. Sci. Rev. **39**, 100336 (2021)
4. Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Schmidt, J., Shah, M.: Decoding brain representations by multimodal learning of neural activity and visual features. IEEE Trans. Pattern Anal. Mach. Intell. **43**(11), 3833–3849 (2020)
5. Nicholson, A.A., Densmore, M., McKinnon, M.C., Neufeld, R.W., Frewen, P.A., Théberge, J., Jetly, R., Richardson, J.D., Lanius, R.A.: Machine learning multivariate pattern analysis predicts classification of posttraumatic stress disorder and its dissociative subtype: a multimodal neuroimaging approach. Psychol. Med. **49**(12), 2049–2059 (2019)
6. Curtindale, L.M., Bahrick, L.E., Lickliter, R., Colombo, J.: Effects of multimodal synchrony on infant attention and heart rate during events with social and nonsocial stimuli. J. Exp. Child Psychol. **178**, 283–294 (2019)
7. Bayoudh, K., Knani, R., Hamdaoui, F., Mtibaa, A.: A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. Vis. Comput., pp. 1–32 (2021)
8. Kaur, P., Malhi, A.K., Pannu, H.S.: Hybrid som based cross-modal retrieval exploiting hebbian learning. Knowl. Based Syst. **239**, 108014 (2022)
9. Zhuang, H., Zhang, J., Liao, F.: A systematic review on application of deep learning in digestive system image processing. Vis. Comput., pp. 1–16 (2021)
10. Karthik, K., Kamath, S.S.: A deep neural network model for content-based medical image retrieval with multi-view classification. Vis. Comput. **37**(7), 1837–1850 (2021)
11. John, L.J.: A review of computer assisted learning in medical undergraduates. J. Pharmacol. Pharmacother. **4**(2), 86–90 (2013)
12. Gabor, A., Popescu, M., Popa-Iovanut, F., Naaji, A.: Telemedicine Technologies, pp. 1–13. Elsevier (2019)
13. Tulsulkar, G., Mishra, N., Thalmann, N.M., Lim, H.E., Lee, M.P., Cheng, S.K.: Can a humanoid social robot stimulate the interactivity of cognitively impaired elderly? a thorough study based on computer vision methods. Vis. Comput. **37**, 3019–3038 (2021)
14. Chowdhuri, S., Pankaj, T., Zipser, K.: In: 2019 IEEE winter conference on applications of computer vision (WACV) (IEEE), pp. 1496–1504 (2019)
15. Ahmad, K., Vrusias, B., Zhu, M.: In: Ninth international conference on information visualisation (IV'05) (IEEE), pp. 268–274 (2005)
16. Guo, Y., Moradi, M.: Cross-modality neural network transform for semi-automatic medical image annotation. US Patent 11,195,313 (2021)
17. Moradi, M., Guo, Y., Gur, Y., Negahdar, M., Syeda-Mahmood, T.: In: International conference on medical image computing and computer-assisted intervention, Springer, pp. 300–307 (2016)
18. Soltani, M.M., Zhu, Z., Hammad, A.: Automated annotation for visual recognition of construction resources using synthetic images. Autom. Constr. **62**, 14–23 (2016)
19. Dutta, A., Gupta, A., Zissermann, A.: Vgg image annotator (via). http://www.robots.ox.ac.uk/vgg/software/via **2** (2016)
20. Zhou, T.H., Wang, L., Ryu, K.H.: Supporting keyword search for image retrieval with integration of probabilistic annotation. Sustainability **7**(5), 6303–6320 (2015)
21. Laaksonen, J., Koskela, M., Oja, E.: Picsom-self-organizing image retrieval with mpeg-7 content descriptors. IEEE Trans. Neural Netw. **13**(4), 841–853 (2002)
22. Viitaniemi, V., Laaksonen, J.: Keyword-detection approach to automatic image annotation (2005)
23. Kaski, S., Honkela, T., Lagus, K., Kohonen, T.: Websom-self-organizing maps of document collections. Neurocomputing **21**(1–3), 101–117 (1998)
24. Mehmood, Z., Mahmood, T., Javid, M.A.: Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine. Appl. Intell. **48**(1), 166–181 (2018)
25. Krishnaswamy Rangarajan, A., Purushothaman, R.: Disease classification in eggplant using pre-trained vgg16 and msvm. Sci. Rep. **10**(1), 1–11 (2020)
26. Rezende, E., Ruppert, G., Carvalho, T., Theophilo, A., Ramos, F., Geus, P.d.: Information technology-new generations, Springer, pp. 51–59 (2018)
27. Ametefe, D.S., Sarnin, S.S., Ali, D.M., Muhammad, Z.Z.: Fingerprint pattern classification using deep transfer learning and data augmentation. Vis. Comput. **39**(4), 1703–1716 (2023)
28. Shah, B., Bhavsar, H.: Depth-restricted convolutional neural network–a model for gujarati food image classification. Vis. Comput., pp. 1–16 (2023)
29. Sharma, V., Tripathi, A.K., Mittal, H., Parmar, A., Soni, A., Amarwal, R.: Weedgan: a novel generative adversarial network for cotton weed identification. Vis. Comput. **9**, 1–7 (2022)
30. Zang, Y., Cao, R., Li, H., Hu, W., Liu, Q.: Mapd: multi-receptive field and attention mechanism for multispectral pedestrian detection. Vis. Comput. **10**, 1–3 (2023)
31. Arulmozhi, P., Abirami, S.: Dshpoolf: deep supervised hashing based on selective pool feature map for image retrieval. Vis. Comput. **37**, 2391–2405 (2021)
32. Ma, J., Wang, T., Li, G., Zhan, Q., Wu, D., Chang, Y., Xue, Y., Zhang, Y., Zuo, J.: Concrete surface roughness measurement method based on edge detection. Vis. Comput. **29**, 1–2 (2023)
33. Paek, S., Sable, C.L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B.H., Chang, S.F., McKeown, K.R.: Integration of visual and text-based approaches for the content labeling and classification of photographs, Acm Sigir, vol. **99**, pp. 15–19. Citeseer (1999)
34. Ibrahim, R.K., Zeebaree, S.R., Jacksi, K., Sadeeq, M.A., Shukur, H.M., Alkhayyat, A.: In: 2021 international conference on advanced computer applications (ACA) (IEEE), pp. 28–33 (2021)
35. Xie, Z., Liu, L., Wu, Y., Li, L., Zhong, L.: Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. IEEE Trans. Serv. Comput. (2021). https://doi.org/10.1109/TSC.2021.3098834
36. Gupta, A., Katarya, R.: Pan-lda: a latent dirichlet allocation based novel feature extraction model for covid-19 data using machine learning. Comput. Biol. Med. **138**, 104920 (2021)
37. Yu, L., Yang, Y., Huang, Z., Wang, P., Song, J., Shen, H.T.: Web video event recognition by semantic analysis from ubiquitous documents. IEEE Trans. Image Process. **25**(12), 5689–5701 (2016)
38. Xu, X.: Artificial intelligence and computer vision, Springer, pp. 165–188 (2017)
39. Li, W., Ma, Z., Deng, L.J., Fan, X., Tian, Y.: Neuron-based spiking transmission and reasoning network for robust image-text retrieval.

IEEE Trans. Circuits Syst. Video Technol. (2022). https://doi.org/10.1109/TCSVT.2022.3233042

40. Wen, K., Tan, Z., Cheng, Q., Chen, C., Gu, X.: Contrastive cross-modal knowledge sharing pre-training for vision-language representation learning and retrieval. arXiv preprint arXiv:2207.00733 (2022)
41. Cheng, Q., Tan, Z., Wen, K., Chen, C., Gu, X.: Semantic pre-alignment and ranking learning with unified framework for cross-modal retrieval. IEEE Trans. Circuits Syst. Video Technol. (2022). https://doi.org/10.1109/TCSVT.2022.3182549
42. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. **43**(1), 59–69 (1982)
43. Kohonen, T.: Essentials of the self-organizing map. Neural Netw. **37**, 52–65 (2013)
44. Pacella, M., Grieco, A., Blaco, M.: On the use of self-organizing map for text clustering in engineering change process analysis: a case study. Comput. Intell. Neurosci. (2016). https://doi.org/10.1155/2016/5139574
45. Li, Z., Qian, Y., Wang, H., Zhou, X., Sheng, G., Jiang, X.: Partial discharge fault diagnosis based on zernike moment and improved bacterial foraging optimization algorithm. Electric Power Syst. Res. **207**, 107854 (2022)
46. Jehangir, S., Khan, S., Khan, S., Nazir, S., Hussain, A.: Zernike moments based handwritten pashto character recognition using linear discriminant analysis. Mehran Univ. Res. J. Eng. Technol. **40**(1), 152–159 (2021)
47. Fredo, A.J., Abilash, R., Femi, R., Mythili, A., Kumar, C.S.: Classification of damages in composite images using zernike moments and support vector machines. Compos. B Eng. **168**, 77–86 (2019)
48. Yang, H., Ni, J., Gao, J., Han, Z., Luan, T.: A novel method for peanut variety identification and classification by improved vgg16. Sci. Rep. **11**(1), 1–17 (2021)
49. Kaur, P., Pannu, H.S., Malhi, A.K.: Comprehensive study of continuous orthogonal moments-a systematic review. ACM Comput. Surv. (CSUR) **52**(4), 1–30 (2019)
50. von F, Z.: Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. Physica **1**(712), 689–704 (1934)
51. Aggarwal, A., Singh, C.: Zernike moments-based gurumukhi character recognition. Appl. Artif. Intell. **30**(5), 429–444 (2016)
52. Teague, M.R.: Image analysis via the general theory of moments∗. J. Opt. Soc. Am. **70**(8), 920–930 (1980). https://doi.org/10.1364/JOSA.70.000920
53. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
54. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
55. Zhang, W., Yoshida, T., Tang, X.: A comparative study of tf* idf, lsi and multi-words for text classification. Expert Syst. Appl. **38**(3), 2758–2765 (2011)
56. Hebb, D.O.: The organization of behavior: a neuropsychological theory. Psychology Press (2005)
57. Oja, E.: Simplified neuron model as a principal component analyzer. J. Math. Biol. **15**(3), 267–273 (1982)
58. Wang, Y., Wu, F., Song, J., Li, X., Zhuang, Y.: In: Proceedings of the 22nd ACM international conference on multimedia, pp. 307–316 (2014)
59. Xie, L., Zhu, L., Chen, G.: Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval. Multimed. Tools Appl. **75**(15), 9185–9204 (2016)
60. Aggarwal, H., Vishwakarma, D.K.: Covariate conscious approach for gait recognition based upon zernike moment invariants. IEEE Trans. Cogn. Dev. Syst. **10**(2), 397–407 (2017)
61. Vlaović, ŽD., Stepanov, B.L., Anđelković, A.S., Rajs, V.M., Čepić, Z.M., Tomić, M.A.: Mapping energy sustainability using the kohonen self-organizing maps-case study. J. Clean. Prod. **412**, 137351 (2023)
62. Rankovic, N., Rankovic, D., Lukic, I., Savic, N., Jovanovic, V.: Unveiling the comorbidities of chronic diseases in serbia using ml algorithms and kohonen self-organizing maps for personalized healthcare frameworks. J. Personal. Med. **13**(7), 1032 (2023)
63. Liu, Z., Feng, S., Zhangsong, A., Han, Y., Cao, R.: Long-term evolution of groundwater hydrochemistry and its influencing factors based on self-organizing map (som). Ecol. Indic. **154**, 110697 (2023)
64. Zhou, G., Miao, F., Tang, Z., Zhou, Y., Luo, Q.: Kohonen neural network and symbiotic-organism search algorithm for intrusion detection of network viruses. Front. Comput. Neurosci. **17**, 1079483 (2023)
65. Slave, A.R., Iojă, I.C., Hossu, C.A., Grădinaru, S.R., Petrior, A.I., Hersperger, A.M.: Assessing public opinion using self-organizing maps. Lessons from urban planning in Romania. Landsc. Urban Plan. **231**, 104641 (2023)
66. Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S.R., Tiede, D., Aryal, J.: Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Remote Sens. **11**(2), 196 (2019)
67. Pannu, H.S., Ahuja, S., Dang, N., Soni, S., Malhi, A.K.: Deep learning based image classification for intestinal hemorrhage. Multimed. Tools Appl. **79**, 21941–21966 (2020)
68. Bekhouche, S., Dornaika, F., Benlamoudi, A., Ouafi, A., Taleb-Ahmed, A.: A comparative study of human facial age estimation: handcrafted features vs. deep features. Multimed. Tools Appl. **79**(35), 26605–26622 (2020)
69. Amato, G., Carrara, F., Falchi, F., Gennaro, C., Lagani, G.: In: International conference on image analysis and processing, Springer, pp. 324–334 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Parminder Kaur is a Postdoctoral Research Associate in the Department of Computer Science at Durham University UK. She completed her PhD from Thapar Institute of Engineering and Technology Patiala, India, and did her Master of Engineering in Computer Science and Engineering. She has an industry experience of 1.5 years as a software engineer in JDA Software, Bangalore, India. Her research interests include Machine Learning, Computer Vision, and Multimodal Information Retrieval.

Avleen Malhi is a Senior Lecturer in Data Science and AI at Bournemouth University UK. She did a postdoc at Aalto University Finland. She also worked as an Assistant Professor at Thapar Institute Patiala, India. Her PhD and Masters are in information security. Her research interests include Machine Learning, Explainable AI, IoT, Ad-Hoc Networks and Information Security.

Husanbir Pannu is a full-time Assistant Professor at Thapar Institute Patiala, India. Earlier, he did a postdoc from the University of Liverpool, UK and Trinity College Dublin, Ireland, PhD from the University of North Texas, USA and MS from California State University Eastbay, USA. His research interests include Machine Learning, Optimisation, and Image Processing.