Int. J. Nonlinear Anal. Appl. 14 (2023) 11, 25–39
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.29216.4088

# Presenting a method based on automatic image annotation techniques for the semantic recovery of images using artificial neural networks

Bita Amirshahi

*Department of Computer Engineering and Information Tehnology, Payame Noor University, Tehran, Iran*

*(Communicated by Seyyed Mohammad Reza Hashemi)*

**Abstract**

With the ever-increasing growth of the Internet and the digital imaging industry, the need to organize and separate images is strongly felt. As a result, image databases with very large sizes were created. In such a situation, there is a strong need for effective tools and methods for image search and recovery. In this proposed method, an automatic image delineation method using an artificial neural network has been presented. At first, by studying reference books and articles related to the basic concepts that are necessary to start working in this field, then by studying the articles of recent years in the desired field, we will examine the strengths and weaknesses of the subject in more detail. Finally, by implementing and testing various ideas that have been expressed in the articles, and with the guidance of the supervisor and applying his ideas, he has tried to cover the weaknesses of other ideas, and finally, a method to improve the accuracy of image indexing and retrieval methods in databases based on the content of images to achieve high efficiency and reduce the semantic gap between low-level features and human perceptual concepts. The proposed color extraction method is simpler, less computationally complex, more accurate and faster. The results of evaluations and comparisons indicate the relative superiority of the proposed method over other methods.

Keywords: Automatic image, Annotation technic, Semantic recovery, Artificial neural networks
2020 MSC: 68T07

## 1 Introduction

Existing application systems for categorizing and retrieving images only categorize them based on the name of the image, and of course, everyone assigns a name to an image based on their perception [17]. Categorizing and indexing images based on content involves extracting a series of visual features from the content of the image, including color, texture, shape, etc., which uses these features to classify the images and restore them based on these features [5]. As a result of personal perception and the passage of time, etc., it has not interfered in naming the image, and as a result of the searches, it has been highly accurate, and of course, in different professions, such image banks are used in different ways [6]. These days, the goal of research in this field is to bring their accuracy closer to the human perceptual system and give importance to high-level features, and restore based on these features [18]. For this purpose and to resolve the semantic gap between retrieval methods based on low-level features and methods that are close to human

*Email address:* b.amirshahi@pnu.ac.ir (Bita Amirshahi)

*Received:* November 2022 *Accepted:* January 2023

intuitive understanding, (due to the existence of a semantic gap between "retrieving images based on content" and "semantic concepts perceived by humans") research has moved towards making connections - bridging this semantic gap - between the low-level properties of the image and the high-level meanings of the image [1]. Communication methods explain semantic distance through automatic image annotation that extracts semantic features using machine learning techniques [8]. This report focuses on recent research in the field of content, based image retrieval methods, and feature extraction as well as a brief study on automatic image annotation methods, and semantic learning methods [21]. In fact, two basic aspects of AIA, which are feature extraction and semantic learning methods, are discussed in this thesis. And finally, by using the features of the proposed method and taking advantage of semantic learning methods, we present our proposed method [7].

With the ever-increasing growth of the Internet and the digital imaging industry, the need to organize and separate images is strongly felt [3]. As a result, image databases with very large sizes were created. In such a situation, there is a strong need for effective tools and methods for image search and recovery [11]. Annotation and image recovery methods allow you to categorize, organize, and filter images based on their contents, and also search and view desired images stored on a computer or network when needed [10]. Now, the higher the accuracy in indexing, categorization and retrieval, the more efficient the systems based on these databases will be, and the problems of non-matching between images will decrease [9]. In this research, by presenting a method based on automatic image annotation techniques, we reduce the semantic distance between image retrieval based on content (color, texture, and shape features) and the semantic concepts perceived by humans, and increase the accuracy of image retrieval and classification.

## 2 Automatic image annotation techniques

### 2.1 Image single-label annotation using Binary categories [12]

In this approach, low-level features are extracted from the image content, and these features are assigned to a classifier and a yes or no answer is received. The output of the category has concepts and meanings that are used for annotation.

#### 2.1.1 Image annotation using support vector machine [4]

A category is needed for each concept, in this case, in the training set, the images that have that concept are positive and the rest of the images are negative. During testing, each category makes a possible decision. The class with the highest probability is the image concept being tested. First, the image must be segmented, and then a number of SVMs (as many concepts) are used. The quality of SVM decreases with increasing concepts.

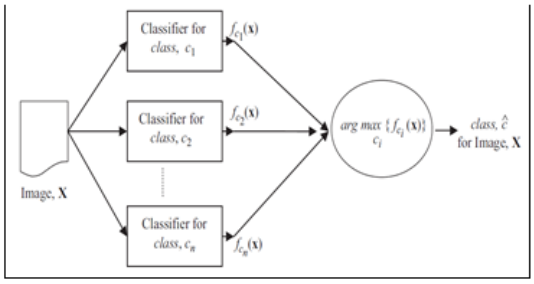



Figure 1: Multi-class classifier using SVM classifiers [12]

To make decision making stronger, several sets can be used, each set containing a number of SVM categories. Each ensemble is trained using a separate subset of the training data. In Figure 2, you can see this method.

#### 2.1.2 Annotation based on rule extraction algorithm for image recovery [2]

Automatic image annotation is used to facilitate searching in large image databases. However, the efficiency of recovery in the existing methods is far from the user's expectation. In this paper, a method for automatic image annotation is presented, this method generates rules using SVM and a decision tree. In order to extract these rules, a set of training regions in the image is collected through image segmentation, image feature extraction, and image discretization.

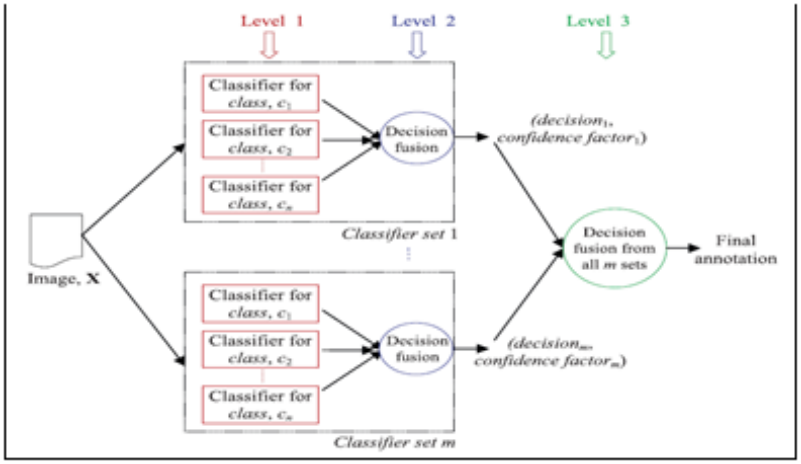



Figure 2: Annotation of images and several sets of SVM [12]

The main goal of this work is to provide a machine-learning method to convert unstructured images into structured text documents. In this system, first, each dataset image is divided into different areas, equivalent to the conceptual objects of the image. In the second step: the visual features of each area, including color, texture, and shape, are extracted, and then the system learns the concepts through discrete values and presents the images in the data set through the trained words. Finally, the system can index and retrieve images. An overview of the presented model is shown in Figure 3.

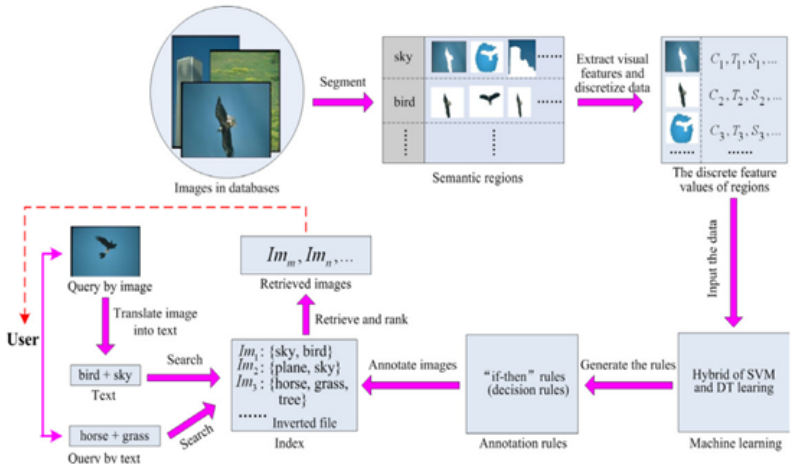



Figure 3: Block diagram of the presented model [16]

### 2.2 Image annotation using artificial neural networks [12]

Curoda and Hagiwara have used four different three-layer neural networks to classify image regions. The number of Noron used in the hidden layer of four networks is 30, 10, 20 and 20 respectively.

Figure 4: shows how the first network classifies an image region into one of the sky, water, and earth classes.

### 2.3 Semantic recovery of images using reverse file [9]

There is as much demand for image management tools as for text search engine tools. Years of research in this field have discovered the existence of a semantic gap between content-based image retrieval systems and the semantic interpretation of the image by humans. As a result, recent research on image retrieval has been oriented towards semantic retrieval. Several semantic recovery models were presented. However, these methods have not yet reached the effectiveness of textual methods in the image field. In this paper, a method is proposed to unify image semantic retrieval with text-based retrieval; this is done by using a reverse file indexing method based on image regions. For this purpose, images are mapped to text documents, and then indexing and retrieval are done through text searches.

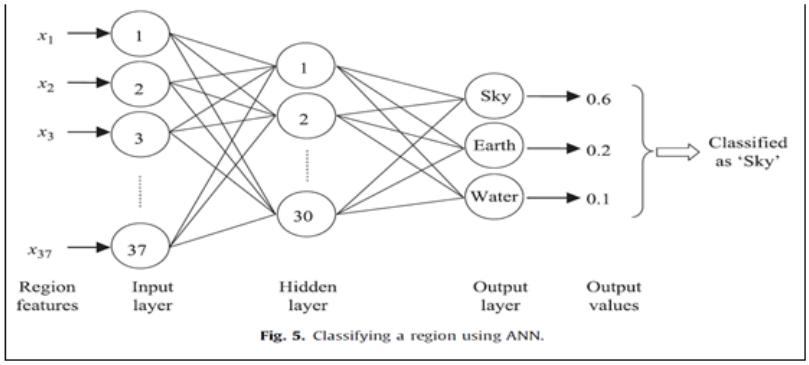



Figure 4: BClassification of a region of the image based on ANN [14]

### 2.3.1 Conceptual learning [9]

The main idea of conceptual learning is to segment the image and extract these regions with the help of training these regions and mapping these images to text documents. This scenario is shown in Figure 5. In the first step, the image is automatically divided into regions using the image segmentation algorithm such as JSEG [19]. These regions are described by their color and texture features (including DCD color feature and Gabor texture features) and are constructed in the order of two banks of visual image features. In fact, each bank contains a set of feature vectors. These banks are actually similar to monolingual dictionaries like the English language dictionary. AVQ algorithm has been used to prepare this dictionary. Each time the dictionary is created, the color property vector of an area replaces the property (password) from the color dictionary. The same thing happens for the texture feature.

When this dictionary is made, it is necessary to establish a connection between the semantic concept and the code word. A set of all these mappings makes a semantic Dictionary, like the English-Persian Dictionary. In this article, the DT decision tree is used to make this mapping [20].

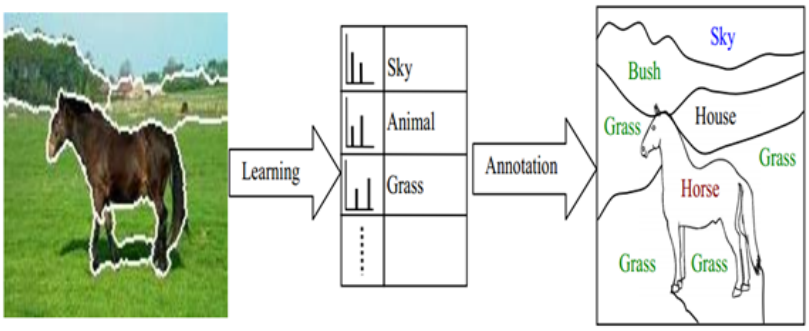



Figure 5: Diagram block of converting images to text documents [13]

Marginalization with the help of DT consists of two steps. The first step is the training step, where the decision tree is trained by labelling regions in the training data set. The decision tree is built based on the basic C4.5 algorithm along with the pruning process. C4.5 [15] is chosen because it has been proven that the feature selection process of C4.5 is much better than ID3. At first, the color and texture features of the image are converted into numerical values. In the next step, the numerical values of the training set are given as DT input for indexing. When the DT is built and trained with the help of the training set, it is used as a semantic dictionary to delineate the uncertain regions of an image in the second step. After annotation, each image in the database is converted into a set of keywords for indexing.

### 2.3.2 Conceptual indexing and recovery using reverse file [9]

Once an image has been mapped to text documents, these images can be indexed and retrieved using the reverse file technique. In this section, we first describe indexing and text recovery using reverse files. Then we describe our presented method which is based on reverse file and image regions to retrieve and index images conceptually.

### 2.3.3 Reverse text files [20]

A reverse text file is a data structure whose documents are indexed to the $< term * document >$ structure, as opposed to the traditional $< document * term >$ structures. Usually, a reverse file is a collection of lists, each list for one expression. To create the reverse file, all text documents are scanned to find candidate keywords and phrases. Suppose the sum of all documents and keywords are N, M in order ($M << N$), the reverse file contains a list for each term. For each expression ($term_j$), the following information is collected and stored:

1. Document repetition frequency : the number of documents that contain the phrase $term_j$.
2. Document ID $doc_j$: Specifies which document contains the phrase $term_j$.
3. Frequency of the phrase $tf_j^i$ : the number of times it appears $term_j$ in document $j$

Table 1: Shows the reverse file for text documents.

| Terms | Document Frequency | Inverted List (Documents & term frequency ) |
|---|---|---|
| $term_1$ | $df_1$ | $< doc_1, tf_1^1 >, < doc_2, tf_1^2 >, \cdots, < doc_{df_1}, tf_1^{df_1} >$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $term_j$ | $df_j$ | $< doc_1, tf_j^1 >, < doc_2, tf_j^2 >, \cdots, < doc_{df_j}, tf_j^{df_j} >$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $term_M$ | $df_M$ | $< doc_1, tf_M^1 >, < doc_2, tf_M^2 >, \cdots, < doc_{df_M}, tf_M^{df_M} >$ |

Table 2: Reverse file structure for text documents [9]

It is clear and evident that if an alternate phrase is repeated in a document, the phrase is very similar to that document. Therefore this $tf_j^i$ ensures that high-frequency terms are given more weight than low-frequency terms. However, if a term appears in multiple documents, the term is not relevant to a particular document and should be given less weight than other terms. It is specified by the inverse frequency $idf_j$ of the document and is defined as follows.

$$idf_j = \log(\frac{N}{df_j}) \tag{1}$$

$idf_j$ and $tf_j^i$ are used to calculate the weight-importance of the statement $term_j$ in document $I$.

$$tw_{i,j} = tf_j^i \times idf_j \tag{2}$$

As a result, document $I$ is introduced with the help of the following feature vector.

$$F_i = <tw_{i,1}, tw_{i,2}, tw_{i,3}, \ldots, tw_{i,M}.> \tag{3}$$

The text document is converted into a feature vector with dimensions M.

$$Q = <q_1, q_2, q_3, \ldots, q_M.> \tag{4}$$

which $q_j$ accepts binary value as follows.

$$q_j = \begin{cases} 1 & if\ term_j\ appears\ in\ the\ query\ text\ Q \\ 0 & otherwise \end{cases} \tag{5}$$

And finally, the similarity between question text Q and document i, $Di$ is calculated as follows.

$$Similarity(D_i, Q) = \frac{\sum_{l \le j \le M}(tw_{i,j} \times q_j)}{\sqrt{\sum_{l \le j \le M} tw_{i,j}^2 \times \sum_{l \le j \le M} q_j}} \tag{6}$$

Then the documents are sorted based on their similarity and returned.

### 2.3.4 Image indexing using reverse file [9]

After the image is mapped to the text document, the image can be indexed and retrieved in reverse file format just like text documents. Anyway, due to the difference between visual documents and textual documents, the concept of the phrase repetition frequency and phrase weight should be reconstructed in the images. Also, additional information such as spatial information of the image should be added to the reverse file to increase the recovery efficiency.

In text documents, each repetition of the phrase in the document gives similar information about the same phrase in other repetitions, so the weight of the phrase is easily determined by the frequency of the repetition of the phrase. The image of each repetition of a concept - phrase - may provide an area of a different size from other repetitions of the same concept. For example, in Figure 6, there are three regions of different dimensions related to the concept of "flower". If the repeat frequency of 2-7 is used, the importance of the tall flower in the center of the image will not be captured, because the other image may have a background smaller than the flower, but not more than 3 areas.

Figure 6: Three areas with different dimensions related to the concept of "flower"[9]

So instead of judging based on term frequency, the regions that map to a term are summed and the "term region" parameter is calculated to indicate the importance of the term "flower".

In addition, the frequency of phrase repetition is another place of conflict between image and text which affects the calculation of the weight of the term [9]. Images consist of a collection of pixels in 2D space, which contains very rich spatial information. For example, as seen in Figure 7, animals usually appear in the center of the image, the sky at the top of the image, and water at the bottom of the image.

Figure 7: An example of locations, sky at the top, water at the bottom, animals at the center [9]

If a region is delineated as an animal region, the retrieval accuracy can be increased provided that this region is

in the center of the image. Therefore, the area of the animal that appears in the center should have a higher weight than the area of the animal that appears in other places of the image. The local weight of each region in the image $I$ is defined as follows.

$$pos\_weight_k^i 2 \times (1 - \frac{d}{d_{max}}) \tag{7}$$

Here, $D$ represents the distance of the area from the center of the normal location in the image and $D_{\max}$ the maximum possible distance. Figure 8 shows an example of how the value of d and $D_{\max}$ is calculated for the animal region, the cloud region, and the water region.

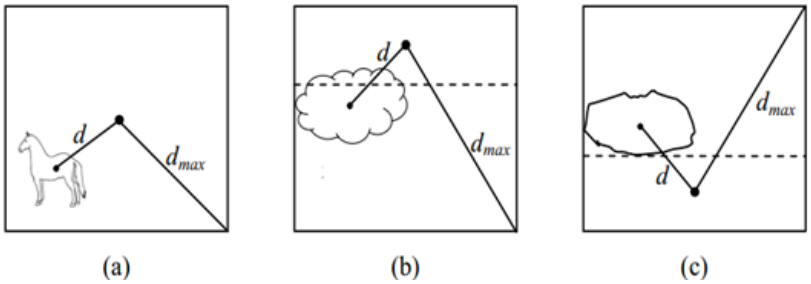



Figure 8: Calculation for a animals b sky c water [9]

The equation 11 stated above ensures that the area farther from the center receives less weight. These parameters are calculated in the same way for other concepts. Table 2 shows how to calculate these parameters for the 3 concepts mentioned here.

Table 2: How to calculate for different concepts [9]

| **Concept of the region** | **Assumption of normal position** | ***d*** | ***$d_{max}$*** |
|---|---|---|---|
| | | *Distance between* | *Distance between* |
| animal, car, fruit, flower | Centre | Image centroid and region centroid | Image centroid and furthest corner |
| firework, sky, bird | Top | Top centre and region centroid | Top centre and furthest corner |
| grass, sand, water | Bottom | Bottom centre and region centroid | Bottom centre and furthest corner |

Important information of Matrix is the proximity of objects in various categories of the image. Certain objects appear simultaneously in the images. [9] For example, as seen in Figure 9, the tiger and the meadow, the birds, the sky, the beach and the sea appear together.

If a region in an image is labelled as a "birds" region, the retrieval accuracy can be improved if this region is associated with "sky" in the same image. Therefore, the weight of the "bird" area increases next to the "sky" area. This idea is introduced using a parameter called the relative weight of the area, which is the frequency of neighboring areas for the area $reg_i$, in image i. [9] For example, if a region is labelled with the term "bird", and two other regions are labelled with the term "sky", the relative weight of the region "bird" is equal to 3. We study our image database and generate a list of neighboring objects. Table 3 shows the number of neighboring objects used in conceptual retrieval.

Because of the information stated above, the reverse file for image documents needs to store more information than text documents. For example, information is stored for each image area. Table 4 shows the reverse file structure for image documents.

Similar to the table, each concept corresponds to the following information set.

1. Document frequency $df_j$ : the number of images that contain one or more regions labeled with the phrase $concept_j, term_j$ .
2. Image $ID_{im_j}$ : Identifies an image that contains one or more labeled regions $term_j$.

Figure 9: Representation of different objects together, tiger and meadow, birds and sky, beach and sea [9]

Table 3: A number of neighboring objects [9]

| Concept Name | Frequently occurring concepts |
|---|---|
| Black Ape | Forest, Brown Grass, Green Grass |
| Black Elephant | Brown Grass, Green Grass, Black Water, Blue Water |
| Brown Horse | Tree, Forest, Brown Grass, Green Grass |
| Building | Blue Sky, Cloudy Sky, Other Sky |
| Butterfly | Tree, Forest, Brown Grass, Green Grass, Flower |
| Camel | Green Grass, Sand |
| Cauliflower | Green Grass, Green Leaf |
| Corn | Green Grass, Green Leaf |
| Deer | Brown Grass, Green Grass |
| Eagle | Blue Sky, Cloudy Sky, Other Sky |
| Eggplant | Green Grass, Green Leaf |
| Fighter Plane | Blue Sky, Cloudy Sky, Other Sky |
| Golden Fish | Black Water, Blue Water |
| Greyhound Dog | Green Grass, Green Leaf, Building, House |
| House | Blue Sky, Cloudy Sky, Other Sky |

3. Expression frequency $tf_j^i$ : the number of areas in image i that have a label $term_j$.
4. Area information list $reginfo_j^i$ : including the number of $tf_j^i$ areas displayed by the following vector.

$$reginfo_j^i = \{(a_1^i, p_1^i, r_1^i), (a_2^i, p_2^i, r_2^i), \ldots, (a_{tf_j^i}^i, p_{tf_j^i}^i, r_{tf_j^i}^i)\} \tag{8}$$

that $a, p, r$ area, $pos\_weight$ , $real\_weight$ represent an image respectively.
The information in the table is used to calculate [9] (It is done when offline indexing).

Table 4: Reverse file structure for image documents [9]

| Term | Image Freq. | Inverted Lists |
|---|---|---|
| $term_1$ | $df_1$ | $\langle im_1, tf_1^1, regInfo_1^1\rangle, \langle im_2, tf_1^2, regInfo_1^2\rangle, \cdots, \langle im_{df_1}, tf_1^{df_1}, regInfo_1^{df_1}\rangle$ |
| ⋮ | ⋮ | ⋮ |
| $term_j$ | $df_j$ | $\langle im_1, tf_j^1, regInfo_j^1\rangle, \langle im_2, tf_j^2, regInfo_j^2\rangle, \cdots, \langle im_{df_j}, tf_j^{df_j}, regInfo_j^{df_j}\rangle$ |
| ⋮ | ⋮ | ⋮ |
| $term_M$ | $df_M$ | $\langle im_1, tf_M^1, regInfo_M^1\rangle, \langle im_2, tf_M^2, regInfo_M^2\rangle, \cdots, \langle im_{df_M}, tf_M^{df_M}, regInfo_M^{df_M}\rangle$ |

$$tw_j^i = idf_j \times \sum_{k=1}^{tf_j^i} (area_k^i \times pos\_weight_k^i \times rel\_weight_k^i) \tag{9}$$

The inverse frequency of the document is also calculated in the same way as equation.

5. The weight of the modified expression is replaced by Equation 7 to extract the feature vector for each image [9].

All images are indexed and retrieved based on these attributes. During retrieval, a query text is given along with one or more keywords. The text of the question is converted into the feature vector of the question according to equation 4 [9]. The similarity between the feature vector of the query image and other images is measured according to equation 10. Finally, the images are ranked based on their similarity.

### 2.4 WEB-based image annotation (using Metadata) [12]

Web is a rich source of visual and textual information. Images on the web are usually with descriptors such as text, URL, HTML, etc. Web information can be used for image annotation or information retrieval.

There are many techniques for annotating images on the Web, most of which use both metadata and visual features. Wang et al. proposed an automatic system for image annotation using web descriptors and content features.

Cluster words with the highest scores are selected for annotation. The advantage of this method is that it does not require a training set. The efficiency of this method depends on web descriptors, which are not very reliable [12].

## 3 Proposed model

The proposed model has two main phases, the training and testing phase, which of course has a phase of extracting the visual features of the image, which is not considered part of the main phase due to its technical details due to the fact that this phase is common to both. We evaluate separately. So, we will continue the content in three general headings: feature extraction, vocabulary network preparation, and expression labeling, and the details of each will be explained below.

### 3.1 Extracting the visual features of the image

#### 3.1.1 Color

In general, color is one of the most important and influential features in image recognition. The human visual system uses color as a tool to find similarity or dissimilarity between images. An accurate image retrieval system usually interprets and uses visual features, especially color, in an extremely effective way. Here we have used the following method to describe the color:

*3.1.1.1. Color similarity criterion*

After extracting the color, to measure the color similarity between two images, it is enough to compare the color percentage of each partition of the first image with the color percentage of each partition of the second image in a

| Annotation method | Pros | Cons |
|---|---|---|
| SVM | Small sample, optimal class boundary, non-linear classification | Single labelling, one class per time, expensive trial and run, sensitive to noisy data, prone to over-fitting |
| ANN | Multiclass outputs, non-linear classification, robust to noisy data, suitable for complex problem | Single labelling, sub-optimal, expensive training, complex and black box classification |
| DT | Intuitive, semantic rules, multiclass outputs, fast, allow missing values, handle both categorical and numerical values | Single labelling, sub-optimal, need pruning, can be unstable |
| Non-parametric | Multi-labelling, model free, fast | Large number of parameters, large sample, sensitive to noisy data |
| Parametric | Multi-labelling, small sample, good approximation of unknown distribution | Predefined distribution, expensive training, approximated boundary |
| Metadata | Use of both textual and visual features | Difficult to relate visual features with textual features, difficult textual feature extraction |

peer-to-peer manner and add the minimum of these two to the total color similarity measurement quantity. Let's add That is, in the form of relationship 10.

$$\text{Color similarity} = \sum_{i=1}^{8} \min(partition\ i^A, partition\ i^B) \tag{10}$$

The closer this value is to one, it means that these two images are more similar in terms of color and vice versa. For example, the color similarity of the above two images is:

$$\text{Color similarity} = 0.5820 + 0.0593 + 0.0049 + 0.0458 + 0.0008 + 0.0025 + 0.0049 + 0.14442 = 0.8444 \quad (10)$$

As a result, the simplicity of the extracted features has led to the design of a simple similarity measure.

### 3.1.2 Texture

Texture analysis involves specifying areas in an image by the relationship of pixels of that image to find repeated patterns in the image. Here, using the concept of texton (derived from Texture Ton), the descriptive texture model tries to determine the concepts of roughness, coarseness, smoothness, smoothness and roughness as spatial changes in pixel intensity [12, 4]. Having information about the local variability of intensity values of a pixel in an image can provide appropriate information about the above concepts, and Texton does this very well. For example, in areas with smooth and soft texture, the range of neighborhood values located around a pixel will be small, while in areas with coarse and rough texture, these values are larger. As a result, obtaining a matrix of values that gives us such information has been very effective in image analysis, because the image texture provides high-level visual information, and its correct analysis leads to a stronger CBIR system. As a result, we decided to use a method for tissue analysis that covers the above issues to an acceptable extent. Here we explain this concept to describe texture.

## 3.2 Preparation of vocabulary network

We go through the following steps to prepare the vocabulary network suitable for the benchmark data (DataSet). That is, in fact, the training stage of the proposed model has several stages, the stage of extracting the visual features of the training images, and the stage of clustering the images:

At first, by selecting the training data, we extract the features stated in the previous paragraph, and extract the feature vector of each image, and finally create a matrix of the feature vector of the training images.

### 3.2.1 Clustering of images

At this stage and after extracting the visual features of the image, all the images are categorized into classes. In order to classify images, it is first necessary to consider a suitable cluster, which is used here using SOM artificial neural network. Before explaining the relationship and usage of this cluster from the extracted features, we introduce this cluster itself.

*3.2.1.1. Clustering of images using artificial neural network*

As seen in Figure 10, the clustering of images takes place in two layers. First, all the images are given to the color feature extraction subsystem and their color features are extracted. These features are given as input to the self-constructing artificial neural network, and the SOM converges by interacting with the weights of its Norons and categorizes the images into four classes by understanding the similarities between the inputs (the number of classes can be four, nine, sixteen, etc.). This number depends on the color distribution in the images as well as their number. In this regard, this network and its related information are stored to be used during retrieval. Next, for all the images in each class, the texture feature vector is extracted and given as input to the self-constructing artificial neural network. SOM also categorizes the images into nine classes based on the received inputs, and performs this operation for each nine classes (the number of classes can be four, nine, sixteen, etc.), this number depends on the texture distribution in the images and their number is dependent. The number nine here is obtained according to the various tests that Corel has performed on the data set. Here, too, this network and its related information are stored to be used during recovery. At this stage, the task of categorizing the images is finished and the information related to the self-generated networks used, saved.

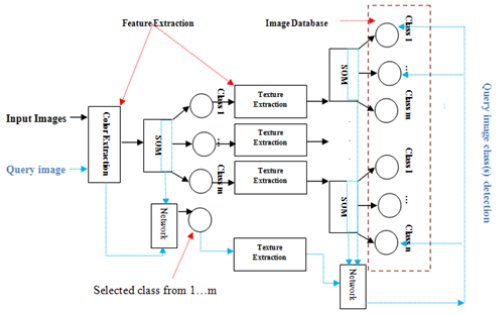



Figure 10: CBIR model architecture

In this proposed plan, the image recovery operation or the clustering of new images in class "I" to which they belong is performed with the help of the feature extraction subsystem and the network information that was saved in the intermediate stages of image clustering. After all the images are indexed in their own classes, to retrieve an image, the color features of the image are extracted and with the help of the network information stored in the color classification stage, the first layer of the query image class is identified. After that, the texture feature vector is extracted and the final class to which the question image belongs is found using the training data and network information that was stored in the classification stage based on the texture. After the question image class is determined, the most similar images to the question images, which are the images with the lowest Euclidean distance from the question images, are returned.

If we call the color and texture feature vector of the query image by $C_Q$ and $T_Q$ and the color and texture feature vector of the images in the database by $C_T$ and $T_T$ respectively, then the Euclidean distance is obtained from equation 11.

$$S(Q,T) = \sqrt{\sum_{i=1...8} (C_Q - C_T)^2 + \sum_{i=1...4} (T_Q - T_T)^2} \tag{11}$$

### 3.3 Labeling

After the above steps, a sequence of labels is calculated and assigned to each image. On the other hand, assuming N labels and T label images will be produced $N^T$, which is not smart and has a lot of errors and will have a lot of calculations. On the other hand, we need a method that has a smart mechanism and less computational complexity

to determine the correct labels. Many methods have been introduced in different sources; below we introduce two of the most famous and widely used ones that have a statistical basis and are used in our proposed model.

# 4 Results

## 4.1 Benchmark data sets

We tested the evaluation basis of the proposed model based on Corel standard image database [16], visTex standard image database [7], and MSRC standard image database.

Standard image database Corel [16], which includes a thousand different images of all kinds of animals such as horses, elephants, dinosaurs, all kinds of cars, all kinds of images of tribal and rural people, sea shores, etc.; and it is one of the most general image databases that are used in most methods for evaluating image retrieval models. The images in this database are in 384*256 dimensions in JPEG format in 10 different categories.

The visTex standard image database [7] was presented in order to provide a large collection of quality texture images for machine vision applications. This collection was created to replace the Brodatz texture database, which is not free for research purposes. VisTex provides images that represent the real world. The MSRC standard image database includes a collection of different images, such as the human face, nature, animals, bicycles, aeroplanes, buildings, clouds, mountains, elephants, landscapes, walls, beaches, sea, boats, crossing, pedestrian, it is river, grass, horse, etc. The images in this database are standard 128x128 images in JPEG format.

## 4.2 Evaluation criteria

To measure the accuracy of the proposed design, two parameters, recall and precision and F-criterion, whose definition is given below, have been used.

Recall: the percentage of retrieved images that match the query image, to the total retrieved images.

Accuracy: the percentage of retrieved images that match the query image, to all images that match the query image in the image database.

F-criterion: a function of recall and precision

The definition of these three parameters is in the form of relations:

$$Recall = \frac{|\{relevnt\ images\} \cap \{retrieved\ images\}|}{relevnt\ images\ indataset} \quad Precision = \frac{|\{relevnt\ images\} \cap \{retrieved\ images\}|}{retrieved\ images}$$
$$F\_measure = \frac{2 * Precision * recall}{Precision + recall} \tag{12}$$

## 4.3 Evaluation results

In order to evaluate the proposed method with other methods, we should have reviewed various articles in this field to use the article that used the CRF model as a classifier and also used the introduced data sets to evaluate their work. In the meantime, we came across the article [20], which we discuss below.

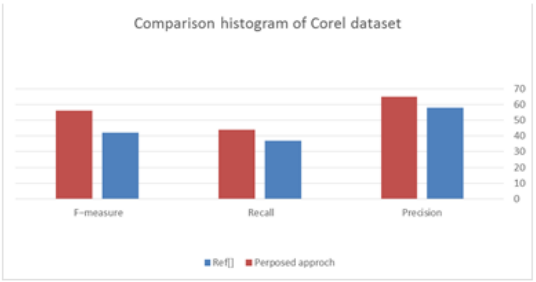
### 4.3.1 Comparison with the method

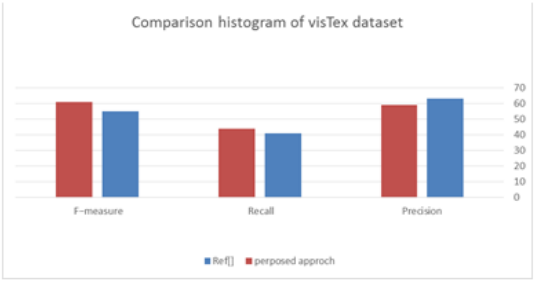
The basis of this method is the presentation and use of feature extraction in one layer, the results of the evaluation and comparison of the proposed method with the method presented in the article are shown in the figure.

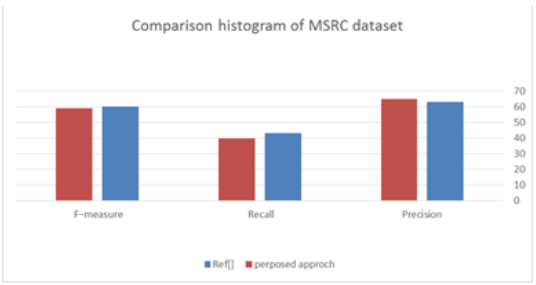
### 4.3.2 Comparison with the method

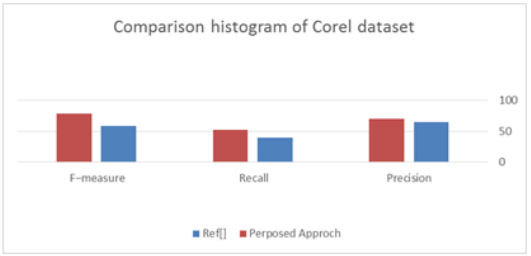
The basis of this method is to present and use feature extraction using image zoning in several layers. The results of the evaluation and comparison of the proposed method with the method presented in the article are shown in the figure.

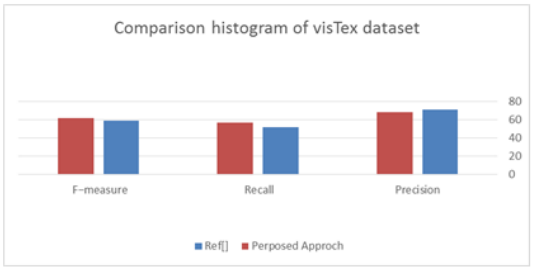
Below is an example of the images annotated by the proposed method from different data sets. In this image, the red lines are the results obtained by the proposed algorithm and the black lines are the actual results of the image.

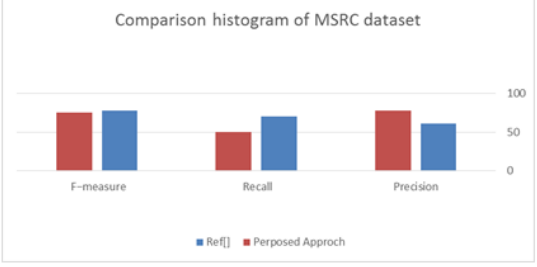

Comparison histogram of Corel dataset
70
60
50
40
30
20
10
0
F-measure
Recall
Precision
Ref[]
Perposed approch


Comparison histogram of visTex dataset
70
60
50
40
30
20
10
0
F-measure
Recall
Precision
Ref[]
perposed approch


Comparison histogram of MSRC dataset
70
60
50
40
30
20
10
0
F-measure
Recall
Precision
Ref[]
perposed approch


Comparison histogram of Corel dataset
100
50
0
F-measure
Recall
Precision
Ref[]
Perposed Approch


Comparison histogram of visTex dataset
80
60
40
20
0
F-measure
Recall
Precision
Ref[]
Perposed Approch


Comparison histogram of MSRC dataset
100
50
0
F-measure
Recall
Precision
Ref[]
Perposed Approch

## 5 Conclusion

In this article, an automatic image annotation method was presented. One of the interesting features of this method is the use of CRF as a model for the optimal mapping of labels according to the extracted features of the images. To evaluate this method, recall criteria, accuracy and F criterion have been used. We also used Corel, visTex, and MSRC standard image databases as benchmark data to evaluate the proposed method. The results of evaluations and comparisons indicate the relative superiority of the proposed method over other methods. The proposed color extraction method is simpler, less computationally complex, more accurate and faster.

The use of the texton concept in extracting the texture, which has not been used in any AIA system, but using this concept borrowed in CBIR, has caused a relative improvement in the proposed AIA system, which ultimately leads to the extraction of features such as energy, contrast, homogeneity and... the description of which is mentioned in the thesis.

For future works, it is suggested to use more expressive visual features, especially the features related to the shapes and objects in the image and to extract the relationship graph between the objects in the image, as well as the use of multi-level clustering, and in fact, to bring processing analysis closer to human analysis, the accuracy of the system Improve auto-tagging.

## References

[1] M.M. Adnan, M.S. M. Rahim, A. Rehman, Z. Mehmood, T. Saba and R.A. Naqvi, *Automatic image annotation based on deep learning models: A systematic review and future challenges*, IEEE Access **9** (2021), 50253–50264.

[2] A. Bahrololoum and H. Nezamabadi-pour, *A multi-expert based framework for automatic image annotation*, Pattern Recogn. **61** (2017), 169–184.

[3] K.A. Kadhim, F. Mohamed and Z.N. Khudhair, *Deep learning: Classification and automated detection earlier of Alzheimer's disease using brain MRI images*, J. Phys. Conf. Ser. **1892** (2021), no. 1.

[4] X. Ke, *Data equilibrium based automatic image annotation by fusing deep model and semantic propagation*, Pattern Recogn. **71** (2017), 60–77.

[5] L. Li, Sh. Tang, L. Deng, Y. Zhang and Q. Tian, *Image caption with global-Local attention*, Thirty-First AAAI Conf. Artific. Intel. (AAAI-17), 2017.

[6] J. Liu and W. Wu, *Automatic image annotation using improved Wasserstein generative adversarial networks*, IAENG Int. J. Comput. Sci. **48** (2021), no. 3, 17.

[7] R. Rad and M. Jamzad, *Image annotation using multi-view non-negative matrix factorization with different number of basis vectors*, J. Visual Commun. Image Represent. **46** (2017), 1–12.

[8] D. Ramachandram and G.W. Taylor, *Deep multimodal learning: A survey on recent advances and trends*, Signal Process. Mag. IEEE **34** (2017), 6–108.

[9] J. Sandra, J. António, C. Mora and A. Almeida, *A novel trademark image retrieval system based on multi-feature extraction and deep networks*, J. Imag. **8** (2022), no. 9.

[10] T. Sun, L. Sun and D.Y. Yeung, *Fine-grained categorization via CNN-based automatic extraction and integration of object-level and part-level features*, Image Vision Comput. **64** (2017), 47–66.

[11] L. Songhao, *Sparse multi-modal topical coding for image annotation*, Neurocomput. **214** (2016), 162–174.

[12] L. Tsochatzidis, *Computer-aided diagnosis of mammographic masses based on a supervised content-based image retrieval approach*, Pattern Recogn. **71** (2017), 106–117.

[13] Zh. Qian, P. Zhong and J. Chen, *Integrating global and local visual features with semantic hierarchies for two-level image annotation*, Neurocomput. **171** (2016), 1167–1174.

[14] D. Tian and Z. Shi, *Automatic image annotation based on Gaussian mixture model considering cross-modal correlations*, J. Visual Commun. Image Represent. **44** (2017), 50–60.

[15] Ch. Tsai, *Bag-of-words representation in image annotation: A review*, Int. Scholar. Res. Notices **2012** (2012).

[16] T. Uricchio, *Automatic image annotation via label transfer in the semantic space*, Pattern Recogn. **71** (2017), 144–157.

[17] A. Vatani, M. Taleby Ahvanooey and M. Rahimi, *Presenting a method based on automatic image annotation techniques for the semantic recovery of images using artificial neural networks*, Int. J. Adv. Comput. Sci. Appl. **9** (2018), no. 3.

[18] R. Wang, Y. Xie, J. Yang, L. Xue, M. Hu and Q. Zhang, *Large scale automatic image annotation based on convolutional neural network*, J. Vis. Commun. Image R. **49** (2017), 213–224.

[19] M. Wang, *Effective automatic image annotation via integrated discriminative and generative models*, Inf. Sci. **262** (2014), 159–171.

[20] J. Zhang, Y. Zhao, D. Li, Zh. Chen and Y. Yuan, *A novel image annotation model based on content representation with multi-layer segmentation*, Neural Comput. Appl. **26** (2015), no. 6, 1407–1422.

[21] S. Zhu, X.Sun and D. Jin, *Multi-view semi-supervised learning for image classification*, Neurocomput. **208** (2016), 136–142.