

Image Annotations using Machine Learning and Features of ID3 Algorithm

D.V.N Harish
M. Tech CSE,
KIET-Engg. College

Y. Srinivas
Dept of IT,
GITAM-University

K.N.V.S.S.K Rajesh
Dept of CSE,
KIET-Engg College,

P. Anuradha
Department of CSE,
Gitam University

ABSTRACT

With the rapid technological growth, availability of digital images in world wide web has increased, since the data levels available in these databases is also increasing enormously, it is very difficult to recognize or retrieve a image of interest instantaneously. Many methodologies have been proposed but still a lot of new technologies are to be proposed in order to retrieve the images of interest basing on the pattern or feature at low cost and with minimum time period. The main disadvantages of the existing methods present in the literature is, the retrieval of images are done with mapping of images using one to one mapping strategy, this method is time consuming, hence to overcome this disadvantage an attempt is made in this paper. This paper focuses on extracting the image if interest across the web in an efficient manner with the pattern and behavior as the key elements.

Key Words

ID3, Image Annotations, Feature set, Machine Learning, Image segmentation, K-Means algorithm

1. INTRODUCTION

With the extensive development in technology, in particular to digital imaging and the internet technologies, the user can easily acquire a huge dataset of images. On the other hand, retrieval of an image is very difficult due to the exponential growth of image database available in the web. It is therefore needed to develop new methodologies which help to retrieve the image or content from the image collections in an efficient and elegant manner with minimum time complexity. Hence, image annotations play a vital role in this regard. The input image is first segmented and the features are extracted and these features are annotated, and after post-processing the labels are identified. Lot of research is available in literature regarding image retrievals and feature vector identification using Bayesian classifiers, SVM's, SVD. But, the main disadvantage in this methods is that Bayesian classifiers are not best suited if the dataset is large[1].SVM and SVD's classify the image on either way for those set of data which lie near the boundaries, hence in either way the main disadvantage of image retrievals using the above foresaid methods is not accurate. Hence in this paper we proposed and develop a new methodology for image annotations using ID3 algorithm. The paper is organized as follows section- 2 deals with a brief discussion about in a segmentation, section-3 deals with K-Means algorithm, section- 4 deals with ID3 algorithm, section- 5 deals with machine learning, section-6 deals with classification with a comparative study section-7 deals with proposed algorithm and experimentation, section-8 deals with testing and results, section-9 deals with conclusions section-10 deals with future extensions. The developed

algorithm is tested on the constructed animal database. The results were compared and tabulated.

2. IMAGE SEGMENTATION

Image segmentation is the prime dissertate for image analysis, it is a process of converting the heterogeneous data into homogenous data [2]. Segmentation can be done in two ways. Non-parametric based segmentation and parametric based segmentation. In non parametric based segmentation, the segmentation process is carried out by using edges, threshold based etc., where as in parametric segmentation, the parameters inside the image are obtained and using these parameters, the segmentation is carried out. Parametric based segmentation is considered to be the efficient when compared to non parametric based methods, because it involves parameters while segmenting [2]. Lot of segmentation algorithms were proposed in literature, but the main disadvantage of segmentation is that the segmentation algorithms differ from application to application [3], another disadvantage is that, to segment the image it is needed to initialize the number of clusters, if the number of clusters are not initialized then the segmentation results differ, hence it is needed for initialization of the number of image components, for this purpose, generally K-Means algorithm is used. The main disadvantage of K-Means algorithm is that if the 'K' is not initialized properly, then it will lead into spherical clusters and finally results into a single cluster, hence to overcome this disadvantage the image segments are first initialized by using a Histogram, this value is passed as the initial estimate for K-Means algorithm.

3. K – MEANS ALGORITHM

The K-Means clustering is a popular approach to segment the image into K-Clusters. The steps to be performed are:

- Step:1. Begin with initial value of k =Number of clusters.
- Step:2. Select the number of clusters k with Initial cluster centroids V_i ; $i=1, 2, \dots, k$. Partition the input pixels into k clusters by assigning each pixel x_j to the closest Cluster centroid V_i using the selected distance measure, e.g. Euclidean distance defined as:
$$d_{ij} = \|x_j - V_i\|, \dots \dots \dots (1)$$
- Step:3. Compute a cluster assignment matrix U representing the partition of the pixels with the binary membership value of the j^{th} pixel to the i^{th} cluster such that:
$$U = [\mu_{ij}] \dots \dots \dots (2)$$

Where,

$$u_{ij} \in \{0, 1\} \forall i, j \dots \dots \dots (3)$$

- Step:4. Recomputed the centroids using
The membership values as
..... (4)
- Step:5. If the cluster centroids or the
assignment Matrix does not change from the
previous iteration, stop else go-to step 3. (5)

4. INDUCTION DECISION TREE (ID3)

This is a iterative model, whose main purpose is to generate all possible decision trees that correctly classify the training set and to select the simplest of them. The number of such trees is finite but very large, so this approach would only be feasible for small induction tasks. In order to overcome this disadvantage we need to construct reasonably good decision tree without much computation. A subset of the training set called the window is chosen at random and a decision tree formed from it; this tree correctly classifies all objects in the window. All other objects in the training set are then classified using the tree. If the tree gives the correct answer for all these objects then it is correct for the entire training set and the process terminates. If not, a selection of the incorrectly classified objects is added to the window and the process continues [4] [5] [6] [7] [8] [9].

5. MACHINE LEARNING

Machine learning is mainly used to recognize complex patterns and make intelligent decisions based on data; the difficulty is with regard to the size of the data (training data) that is used. The image is segmented into basic blocks, which classifies according to the type of content, based on color, texture and behavior. A lot of research is projected towards image annotation based on (support vector machines) SVM and Bayesian networks.

However the main disadvantage of classifying the data using SVM is that it classifies the data is marginally and the main disadvantage of SVM classifier is the data which is nearest to the line will be classified in either way. Hence, there is every possibility of misclassification where as in Bayesian classification, it needs a very complex data set and time taken to classify the data is marginally high. Hence, in a proposed method we develop a methodology for image annotation based on ID3 algorithm the results obtained from this method are tabulated.

ID3 algorithm based on the following two suppositions:

- (1)The class probability that a correct decision tree classify to random example set is consistent with the probability of positive-example and negative-example in vector space E.

$$\frac{p}{p+n} \dots \dots \dots (6)$$

A is the event, Where p is the positive example i.e., the possibility of happening and n is the negative example i.e., possibility of not happening

- (2) The information entropy needed for a decision making, and is calculated as follows

$$-\frac{p}{p+n} \log \frac{p}{p+n} - \frac{n}{p+n} \log \frac{n}{p+n} \dots \dots (7)$$

- (3) The information gained is calculated using
Gain =I (p,n)-E(A).....(8)

ID3 algorithm chooses the attribute with maximum Gain (A) as the root node, which means the attributes with the minimum E (A).

6. CLASSIFICATION

To classify the feature set obtained after clustering using the k- means algorithm, ID3 algorithm is used, the main purpose of ID3 algorithm is to find all the possibilities in the decision trees which enables as to classify the training set more accurately and exactly. The experimentation is carried out by taking a sample animal database of size 20 with 4 different set of animals namely Mammals, Birds, Aquatic, Amphibians, this data set is given as an input to the K Means Algorithm and the individual clusters. We then train the dataset with the three different concepts namely color, texture, behavior.

The feature set can be of two types either a local feature set or a global feature set, we have designed a questioner for the color discrimination, texture discrimination and also behavior discrimination. To classify the color we have considered the animal color into three types i.e., black, brown and yellow. The texture is considered while classifying are three types i.e., Plain, Stripes and Spotted, to identify the Behavior we have considered the animal behaviors into two types i.e., carnivores and herbivores.

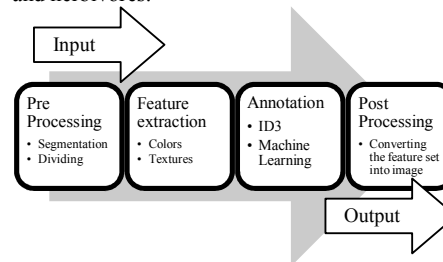


Figure 1: Block diagram of the image annotation system

An intuitive answer to this question is to segment the image into regions, cluster similar regions and then use the regions as a vocabulary hence we obtained a good vocabulary.

7. PROPOSED ALGORITHM

1. Consider the image database
2. Apply the k means algorithm to identify the cluster
3. Obtain the features color, texture and behavior
4. Apply the ID3 algorithm for classification of images basing on the feature set deriving on step 2
5. The step 4 is continued until all the images are classified, all the image in the data set using the features.

The results of the experimentation using the above algorithm is carried out in the image data set and is presented in table 1 the output images obtained are presented in figure-2 represents the data set figure-3 and figure-4 represents the process of classification is represented in figure-5.



Figure-2 representation of Data Set



Figure-3 Retrieving the images using classification models

Sl.No	Name of the image	Feature sets			Output image in database
		Color	Texture	Behavior	
1	Wild	Black	Stripes	Carnivores	White tiger
				Herbivores	Zebra
				Carnivores	White cheetha
		Spotted	Plain	Herbivores	White horse
				Carnivores	Panther
				Herbivores	Bull
		Brown	Stripes	Carnivores	Long lizard
				Herbivores	Squirrel
				Carnivores	Anaconda
Spotted	Herbivores		Indian cow		
	Carnivores		Bear		
	Herbivores		Australian horse		
Yellow	Stripes	Carnivores	Tiger		
		Herbivores	African deer		
		Carnivores	Hyena		
	Spotted	Herbivores	Giraffe		
		Carnivores	Lion		
		Herbivores	Donkey		
4	Bird	Black	Stripes	Carnivores	Owl
				Herbivores	Sparrow
				Carnivores	Vulture
		Spotted	Plain	Herbivores	Parrot
				Carnivores	Eagle
				Herbivores	Peacock
		Brown	Stripes	Carnivores	Vampire bat
				Herbivores	Brown bull
				Carnivores	African eagle
Spotted	Plain		Herbivores	Brown pelican	
			Carnivores	Indian vultures	
			Herbivores	Gyrfalcon	
Yellow	Stripes	Carnivores	Australian eagle		
		Herbivores	Pigeon		
		Carnivores	American owl		
	Spotted	Plain	Herbivores	Quail	
			Carnivores	Bad bird	
			Herbivores	Woodpecker	
7	Amphibians	Black	Stripes	Carnivores	Lizard
				Herbivores	Frog
				Carnivores	Dragon
		Spotted	Plain	Herbivores	Turtle
				Carnivores	Snake
				Herbivores	Sea turtle
		Brown	Stripes	Carnivores	Gladiator
				Herbivores	Seal
				Carnivores	Crocodile
Spotted	Plain		Herbivores	Penguin	
			Carnivores	Godzilla	
			Herbivores	Ortonil	
Yellow	Stripes	Carnivores	Attractus schach		
		Herbivores	Placa		
		Carnivores	Iphasa elegans		
	Spotted	Plain	Herbivores	Femorialis	
			Carnivores	Cetnophryne geayi	
			Herbivores	Magnusoni	
10	Aquatic	Black	Stripes	Carnivores	Giant isopod
				Herbivores	Fish
				Carnivores	Octopus
		Spotted	Plain	Herbivores	Gelly Fish
				Carnivores	Lazydisis
				Herbivores	Sea horse
		Brown	Stripes	Carnivores	Sea elephant
				Herbivores	Sea horse
				Carnivores	Sea lion
Spotted	Plain		Herbivores	Anger fish	
			Carnivores	Brown skate	
			Herbivores	Skate	
Yellow	Stripes	Carnivores	Killer whale		
		Herbivores	Whale		
		Carnivores	Golden fish		
	Spotted	Plain	Herbivores	Yellow shark	
			Carnivores	Isopod	
			Herbivores	Good fish	

Table: 1 The classification and differentiation between images using the image data base after and before applying the algorithm

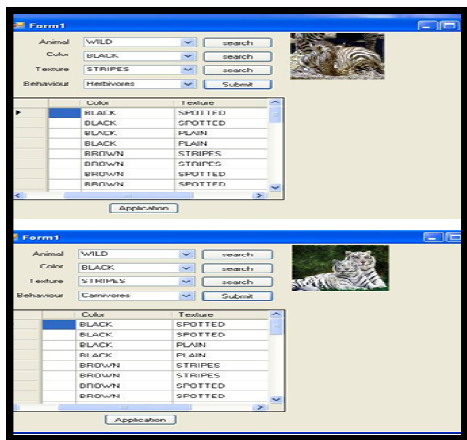


Figure-4: Showing the difference between two sets of images in behavior

The output results are evaluated using image fidelity which consists of two criteria's, the objective criteria and the subjective criteria where the set of input images and the retrieved images are showed to the viewers the outcome of the result shown is very good classification rate

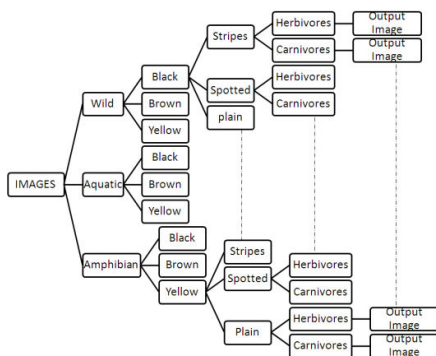


Figure-5 classification in diagrammatic manner

8. TEST RESULTS

In order to demonstrate our model we have considered a data set of 72 images consisting of animals, birds, aquatic and amphibians. We have considered the features based on color, texture and behavior. We have taken into consideration two types of behavior classes mainly flesh eating (carnivores) and grass eating (herbivores). The database is first segmented basing on the query of interest i.e., either to retrieve animals, birds, aquatic or amphibians and the k means algorithm is utilized for this purpose. In order to classify or in order to retrieve the query of interest exactly we have used the ID3 algorithm with color, texture and behavior as the patterns for the efficient retrieval. The query processed and retrieved images are presented in figure-2 and figure-3 the output results from these two figures clearly show that query of interest is retrieve accurately. The efficiency of the algorithm is clearly seen from the above figures and our algorithm works more exactly and accurately since we are considering multiple patterns to retrieve a query of interest.

9. CONCLUSION

A novel algorithm for image annotation using ID3 is presented since the input image is un supervised data K means algorithm is used for clustering so that the classification technique is applied the feature vector in our mythology consists of both local and global feature sets using IF-THEN loops to identify the feature set we have used three feature sets for extraction of image accurately i.e., using color texture behavior the output obtained from the mythology consists of a image with the above feature set the performance of the developed algorithm is evaluated by the image fidelity this go for further extension is also presented

10. FUTURE EXTENSIONS

We want to enhance the development of our algorithm using the concept of boundary value problem where we want to have much more effective classification and retrieval using the technique Boundary Value Problems(BVP) , this is Another way to produce the annotation of the image is boundary value problem in electrical circuits mainly to detect the boundaries of the circuit, first the pixels should be calculated for the image next the adjacent pixels should be connected with a line, and the adjacent lines should be connected to get the shape of the images

11. APPENDIX

Partial Pseudo Code for ID3

```

If the animal is in jungle
    Then go to wild animal ()
If the animal is in air
    Then go to bird ()
If the animal in water
    Then go to aquatic ()
If the animal is in both water and land
    Then go to amphibian ()

Wild animal ()

If image color code is#000000
    Then go to Black
If image color code is#660000
    Then go to Brown
If image color code is#FFFF00
    Then go to Yellow
{
Black ()
{
    If image is flesh eating
        Then go to Carnivores
    If image is Grass eating
        Then go to Herbivores
}
Carnivores ()
{
    If the image having stripes
        Then it is a jaguar
    If the image having plain
        Then it is a bear
    If the image having spotted
        Then it is a killer whale
}
Herbivores ()
{

```

