Combination of Complementary Features for Automatic Image Annotation

Rekhil M Kumar
Dept. of Computer and Information Science
College of Engineering, Poonjar
Kottayam, Kerala, India

Sreekumar K
Dept. of Computer and Information Science
College of Engineering, Poonjar
Kottayam, Kerala, India

ABSTRACT
Image annotation is a method for representing an image with a suitable keyword closer to its semantic concept. Automatically assigning relevant text keywords to image is an important problem. Many algorithms and combination of different features have been proposed in the past and achieved good performance. Efforts have focused upon many other fields and some predefined set of features in the area of Automatic image annotation. But properties of features and their complementing combinations have not been well investigated. In this paper the performance of different feature combinations are compared, and find out the one which outperforms the other combinations by applying the Fuzzy K-nearest neighbor algorithm as the classification method.

General Terms
Computer vision, Image Processing.

Keywords
Automatic Image Annotation (AIA), Feature Extraction, Binary Descriptor, color Descriptors, Texture Descriptors.

1. INTRODUCTION
Now a day, there is a dramatic increment in the field of image capturing devices and digital images. So the problem of finding an image from a large image set is significant. A number of search engines retrieve images based on a text keyword without using content information. Automatic image annotation is proposed to overcome these kinds of issues. AIA will automatically assign closest text keywords to any given image, reflecting its content properly.

The semantic gap between high level visual aspects and low level visual features can be minimized with the help of Automatic Image Annotation. AIA and its applications have a lot of to do with Social networking websites and image databases [7].

Some of the existing methods in AIA have the problem related with pre-selection of features without considering their properties. But a well investigation about the combination of different feature descriptors based on the properties will improve the image annotation task. In effect predefined feature sets do not contribute to the performance of annotation positively.

This paper presents a perfect combination of feature descriptors regarding color, texture and an algorithm for object detection with the help of a classification algorithm-FK-NN [1] for classifying images automatically. Comparison between different image feature descriptors based on color, texture and frequency domain features have already been performed to solve the feature related problems in the image annotation.

2. RELATED WORK
The former approaches proposed for annotating images were Image tagging and content based image retrieval. Image Tagging can be considered as a process of assigning a metadata to a piece of information in an image. Even if it increases the amount of retrievals, the individual tags may disjoint, irrelevant and confusing. This will lead Image Tagging lack precision. When comes to Content Based Image Retrieval (CBIR) [2], It only uses the visual aspects of an image such as color, texture and shape for image representation. Thus it faces the difficulty in locating desired image from a large collection of images .Along with it faces the problem of matching with human visual system.

AIA outperforms the two methods by its high performance and the ability to perform effective manipulation for exponentially growing photo collection. The probabilistic modeling methods and the classification methods are the two recent developments in the area of AIA. In the probabilistic modeling the correlation between images and keywords is represented by a relevance model. On the other hand, the (classification) model trains a separate classifier from visual features for each tag. These classifiers are used to predict particular tags for test image samples.

At high level perspective a feature is used to represent an image as a metric or some quantifiable value. Features related to color, texture, shape are usually used to represent an image .The first step is to detect interest points in the image having the property of repeatability, means the ability to detect the same physical interest points under different viewing conditions, followed by the description calculation of the interest points. The features selected to be unique i.e. if similar point is being described in two or more images then that point should have similar description and it should be of proper dimensions, a large descriptor will makes the computation longer. But if the descriptor is small then it may discard some useful information.

One of the extensively used feature for annotating images automatically is the color feature For color feature description a number of algorithms are existing, such as color histograms (RGB&HSV)[3] ,color moments [4], color coherent vector(CCV) [5], color structure Descriptor (CSD),color layout descriptor(CLD) .Out of these SCD is selected after the experiment due to it allowing a trade-off between accuracy and speed [6].

The very next feature, played a significant role in Image Annotation is Histogram Of Oriented Gradients [8] (HOG). Object detection, especially human perception is an important problem in many cases. To overcome the problem of object detection there is a need of an object detection algorithm. Normalized histogram of oriented gradients (HOG) is used here to solve the issue. HOG significantly outperforms the other feature sets for this purpose.

The texture feature is the next perfect feature for annotation procedure. Texture usually aims to find a unique way of representing the characteristics of textures and give a definition to them in a simpler way. Gray Level Co-occurrence Matrix (GLCM)[9] possesses high efficiency in order to extract second order statistical texture features.

After the detailed experiments with each of these features and their combinations, the combination of SCD and HOG produce better results. By the addition of GLCM, the result has a slight improvement in terms of accuracy.

3. PROPOSED SYSTEM

Training and Annotation are the two main phases involved in the proposed system. Training phase starts with the SCD Color feature extraction, HOG Feature extraction and GLCM texture feature extraction of the images to be trained. After the extraction step completed, the next step is to concatenating the feature sets for each class and it result in a model descriptor for each class. Thus after the training phase, each of the class will be represented with a unique feature vector.

Annotation phase starts with the same feature extraction procedure as did in the training phase for the images to be tested.

Then the image is classified according to the fuzzy- knn classification algorithm. And finally for the annotation purpose, the model descriptor from the training phase is used. The proposed architecture is shown in fig (1).

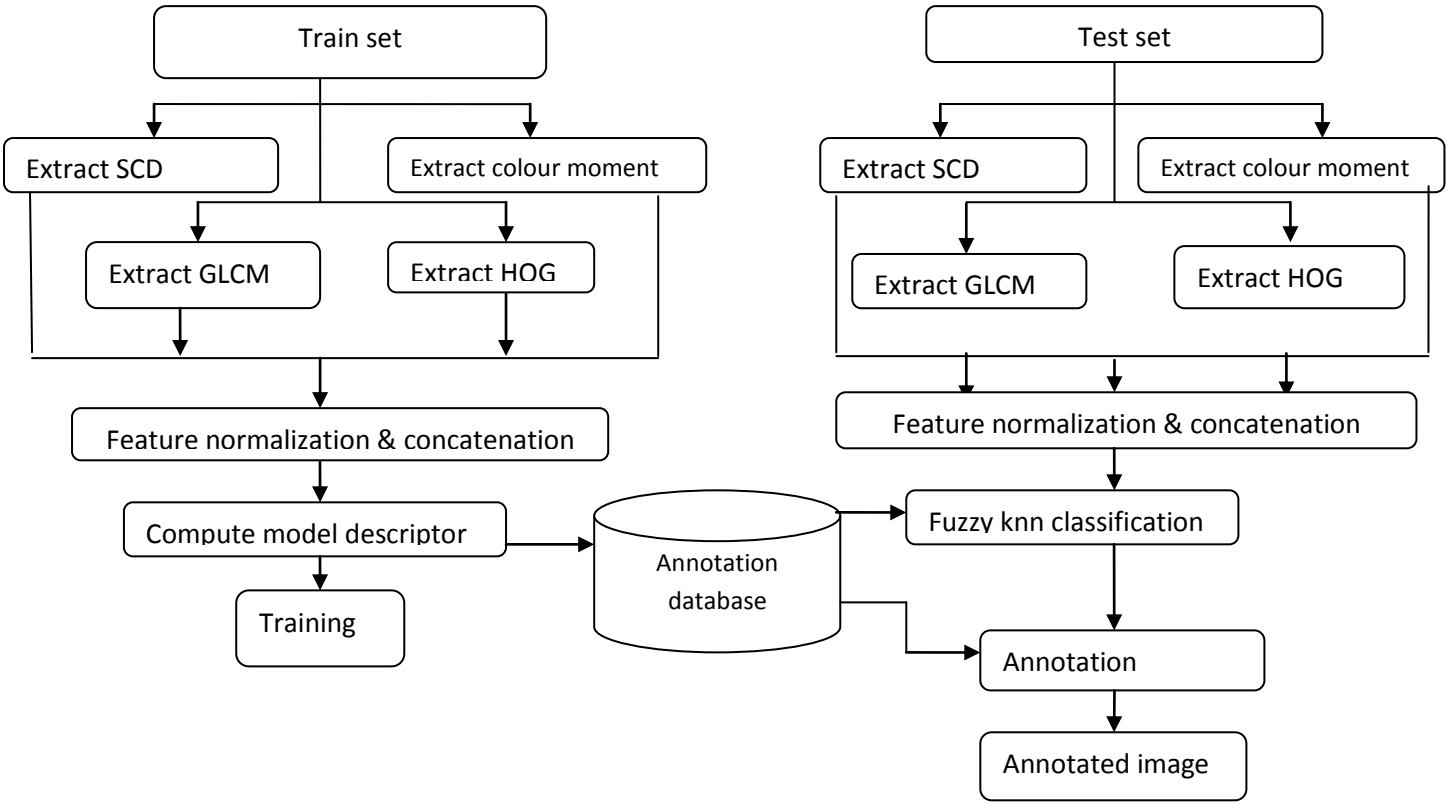

Fig 1: Architecture of the proposed system

3.1 Feature Extraction

The accuracy of any image annotation task will completely depend upon the feature set selection process. The features regarding color, texture and shape are mainly used for annotation purpose. Here the implementation starts with extraction of SCD color feature, HOG feature, and GLCM texture feature along with color moments.

3.1.1 SCD Color Feature

The Scalable Color Descriptor gives a compact representation of an image by using the HSV color space, still maintaining the accuracy of classification. In this algorithm the amount of bins/histogram can be set to a number of different ways. It allows scalable description representation through Haar-Transform co-efficient encoding. In total SCD is attractive due to it offering scalability in terms of number of bins by varying number of coefficients used.

The first step for SCD feature extraction is to find out the histogram values of HSV color space. The colors of an HSV image are uniformly quantized into 256 bins. In order to calculate a histogram with half number of bins, summing up pairs of adjacent bins has to be performed. After the summation of every adjacent Hue bin value, a 128 bin histogram representation constitute of 8-levels in H, 4-levels in S, 4-levels in V. If this process is repeated for 64, 32, 16 sum coefficients from the Haar representation, it is similar to histograms of 64,32,and 16 bins. Thus here we get a 384 dimensional feature vector for representing SCD color feature.

The histogram values are extracted, normalized and non-linearly mapped into a 4-bit integer representation. And finally Haar transform is applied to this integer value across all the bins. A sum operation and difference operations are the resolution levels with large number of bins can be expressed by the high-pass (difference) coefficients.

Comparisons of different size representations in the SCD allow the application of a coarse-to-fine procedure, achieving significant speed in similarity matching over a large dataset.

3.1.2 HOG (Histogram of Oriented Gradients)

SCD color feature is itself not outstanding for the class "AFRICANS "in our dataset 'COREL 1000',as it contains images of humans clearly. Thus HOG is the robust feature to solve the issue related with human/object detection. Concatenating SCD with HOG gives better results than what SCD alone performs.

The HOG strictly based on an evaluation of local histograms (normalized) of image gradient orientations in a dense grid. It will reveal the edge characteristics of an image which in turn can describe the local shape of the image.

The algorithm will start by dividing the image into small connected cells. And for each cell, for the pixels within the cell calculate the histogram of gradient directions or edge orientations. The next step is to divide each cell into a number of angular bins according to the edge orientations. The pixel from each cell contributes weighted gradients to the corresponding angular bin.

One of the important aspects related with HOG is 'blocks'. The adjacent cells are grouped together to form spatial regions, which are called as blocks. The fundamental concept behind the normalization of histograms is the grouping of cells into blocks. Finally a block histogram is formed from the normalized group of histograms. The set of block histograms represents a complete descriptor of size 81, with 3× 3 cells and 9 bins per histogram .

3.1.3 GLCM Texture Feature

The Gray Level Co-occurrence Matrix is one of the texture feature extraction algorithm by which we can get information about the structural arrangement of the object surface. Second order statically texture features from an image are extracted by GLCM.

GLCM matrix represents the number of gray levels of an image. Each of the element (p(i,j)) in the GLCM matrix indicates how often a pixel with gray level (intensity) value i occurs in specific relationship with another pixel with value j. The spatial relationship indicates the direction of pixel of interest to the nearest pixel we consider. The co-occurrence matrix can be formed from the following equation.

$$\begin{cases} 1: \textit{If } I(x,y)=I \textit{ and } I(x+\Delta_x,\Delta_y)=j \\ : \; P(I,j)= \Sigma_{x=1}\Sigma_y \\ \quad \text{Otherwise} \\ 0 \end{cases} \tag{1}$$

Here the value (Δx,Δy) denotes the offset i.e., distance between pixel of interest and its neighbor. The values of offset make the GLCM sensitive to rotation. The four main offset values are,

- P-horizontal(0 degree) : (0,Δ)
- P-right diagonal(45 degree) : (-Δ,Δ)
- P-vertical(90 degree) : (-Δ,0)
- P-left diagonal(135 degree) : (-Δ,-Δ)

After the GLCM has been created, we can derive several statistics from them by using formulas. The important ones are listed as,

3.1.3.1 Energy(Angular second moment)

It measures the textural uniformity. It can also be referred as the sum of squares if entries in a GLCM matrix.

$$\text{Energy}= \Sigma_i \Sigma_j \, P_{ij}^{\,2} \tag{2}$$

3.1.3.2 Entropy

It measures the complexity of an image. If an image has higher complexity then it would have more entropy value.

$$\text{Entropy}=\Sigma_i \Sigma_j \, P_{ij} \log_2 P_{ij} \tag{3}$$

3.1.3.3 Contrast

It measures the difference between the highest and lowest values of a contiguous set of pixels, in turn it can specify the local variations present in an image.

$$\text{Contrast}= \Sigma_i \Sigma_j \, (i-j)^2 \, P_{ij} \tag{4}$$

3.1.3.4 Variance

This is used to measure the heterogeneity, and will increases when the gray level value deviates from its mean.

$$\text{Variance} = \Sigma_i \Sigma_j (i-\mu)^2 P_{ij} \quad (\mu = \text{mean of } P_{ij}) \tag{5}$$

3.1.3.5 Homogeneity (Inverse difference moment)

It assumes larger values when there exist smaller gray tone differences in pair elements, thus gives the value of homogeneity.

$$\text{Homogeneity} = \Sigma i \, \Sigma j \, (1/(1+(i-j)^2)) \, P_{ij} \tag{6}$$

3.1.3.6 Correlation

It gives gray tone linear dependencies in an image.

$$\text{Correlation} = (\Sigma_i \Sigma j \, (ij) P_{ij} - \mu_x \, \mu_y)/\sigma_x \, \sigma_y \tag{7}$$

The rest of the textural features namely sum average, sum entropy, sum variance, difference variance, difference entropy, maximum correlation coefficient and information measures of correlation are secondary, derived from the above six important ones. Here we extract the important 11 features, which in turn give a 44 dimensional feature vector for GLCM. When extracting the features of an image with GLCM approach, at the time of RGB to GRAY level conversion the image compression time can be greatly reduced.

3.1.4 Colour Moments

Colour moments are measures that characterize colour distribution in an image. Colour moments are mainly used to compare how similar two images are based on colour. Usually one image is compared to a database of digital images with pre-computed features in order to find and retrieve a similar image. Each comparison between images results in a similarity score, and the lower this score is the more identical the two images are supposed to be.

Colour moments are scaling and rotation invariant. It is usually the case that only the first three colour moments are used as features in image retrieval applications as most of the colour distribution information is contained in the low-order moments. Colour moments can be computed for any colour model. Three colour moments are computed per channel (e.g. 9 moments if the colour model is RGB and 12 moments if the colour model is CMYK). Here we consider RGB colour model for our implementation, Mean and standard deviation is calculated for each of the three channels R G and B. Thus we get a descriptor of size 6 here. Computing colour moments is done in the same way as computing moments of a probability distribution.

On combining GLCM along with the combination of SCD, HOG and color moments, the accuracy of the proposed system increases slightly, ie, the number of images classified to their corresponding classes incremented by 10 percentages.

3.2 Training phase

The training dataset comprises with a total of 'n' images, and is clustered into 'k' images in each of the 10 classes .A 509 dimension feature vector is extracted for each of the 'k' image in all the classes, and a fused feature vector of size 509 is calculated for representing each class. This is the model descriptor for each class, which we further used in annotation phase.

3.3 Annotation phase

The annotation phase proceeds with the calculation of the feature vector of size 515 for all the test images. These will same as that of the feature vector calculation done in the previous stage. The next step is to create a description matrix with the feature vector representing all images in the trained dataset. Here each column stands for 515 dimensional feature vectors for each of the image .The training matrix which we have already done consists of a 515 dimensional feature vector for each class, i.e. each column will represent a particular class and its 509 dimensional feature vector. Fuzzy k-nearest neighbor algorithm is used as a classifier. It assigns a class membership value for each of the input image . Euclidian distance is used to measure the similarity between the vectors of image to be tested to the vectors of images in the training set. The Euclidian distance is measured by the equation,

$$d(p,q) = \sqrt{\Sigma^n_{i=1} (q_i - p_i)^2} \tag{8}$$

And class membership value prediction will be done by u the equation,

$$U_j(x) = \left(\Sigma^k_{j=1} u_{ij}^{(1/\|x-x_j\|^{2/(m-1)})}\right) \Big/ \left(\Sigma^k_{j=1}{}^{(1/\|x-x_j\|^{2/(m-1)})}\right) \tag{9}$$

According to the class numbers given by the fuzzy k-nn we can retrieve the corresponding class names, and able to annotate each image.

4. RESULTS AND PERFORMANCE EVALUATION

The COREL 1000 dataset was used for the implementation and experiments. The dataset contains total 1000 images which are clustered into 10 classes, which are, Africans, Beach, Building, Bus, Dinosaurs, Elephant, Flower, Food, Horse, Mountain. The entire images have undergone training procedure. Annotation was done for around 775 images from 1000. It results no repetition of annotation for the same image. The classification was done by fuzzy k-nn algorithm by measuring Euclidian distance.

The system for automatic image annotation was implemented with a number of feature combinations regarding colour, texture, shape and frequency domain features in order to find the one combination which outperforms the other ones. Some of the combinations with comparable results and their performance are given below,

- colour histogram + Maximally Stable External Regions(MSER) + Histogram of Oriented Gradients(HOG)
- Gray Level Co-occurrence Matrix(GLCM) + HOG + SCD + Fast Retina Key Point(FREAK)
- HOG + GLCM + HSV HISTOGRAM
- Binary Robust Invariant Scalable Key points(BRISK) + GLCM + SCD

The performance of these combinations was compared in terms of their precision, recall and accuracy, which can be figured out from the comparison matrix 1 and comparison matrix 2.

Table 1. Comparison matrix 1

Feature	HSV + MSER + HOG			FREAK + GLCM + SCD + HOG		
Class Name	**P**	**R**	**A**	**P**	**R**	**A**
Africans	0.73	0.73	0.95	0.79	0.79	0.96
Beach	0.71	0.71	0.94	0.79	0.79	0.96
Buildings	0.60	0.60	0.92	0.77	0.73	0.95
Bus	0.91	0.91	0.98	0.96	0.96	0.99
Dinosaurs	0.93	0.88	0.98	0.98	0.98	0.99
Elephant	0.60	0.50	0.92	0.75	0.55	0.94
Flower	0.91	0.84	0.98	0.87	0.87	0.97
Food	0.73	0.73	0.5	0.72	0.71	0.94
Horse	0.75	0.75	0.95	0.82	0.82	0.96
Mountain	0.77	0.69	0.95	0.71	0.71	0.94

The first combination with HSV colour histogram, maximally stable external regions and HOG gives an average accuracy of 95.2 percent. The combination has been tried out because of it contributes one colour feature, one frequency domain feature along with a feature for human detection.

Table 2. Comparison matrix 1

Feature	HOG + GLCM + HSV			GLCM + SCD + BRISK		
Class Name	P	R	A	P	R	A
Africans	0.77	0.77	0.95	0.76	0.76	0.95
Beach	0.71	0.71	0.94	0.70	0.70	0.94
Buildings	0.75	0.61	0.94	0.65	0.65	0.93
Bus	0.93	0.93	0.99	0.80	0.80	0.96
Dinosaurs	0.92	0.88	0.98	0.98	0.98	0.99
Elephant	0.58	0.46	0.91	0.68	0.49	0.93
Flower	0.94	0.88	0.98	0.00	0.00	0.90
Food	0.66	0.66	0.93	0.63	0.63	0.93
Horse	0.78	0.78	0.96	0.73	0.73	0.95
Mountain	0.93	0.71	0.97	0.71	0.71	0.94

Here MSER is a 64 dimensional feature vector HOG is of 81 and hsv of 10, a total of 155 dimensional feature vectors is used to represent images. Though the minimum descriptor size reduces processing time, which does not provide accuracy comparable with existing ones. When comparing next combination of HOG + GLCM + HSV histogram, our motivation was just find out how a texture feature would work with the combination of HOG and HSV instead of MSER. It results better accuracy than the previous, and is about 95.5 percent. This makes the combination more satisfactory. As GLCM is a 44-D feature vector, this combination produces a 135-D feature vector for images. For the class flower, it results an accuracy similar to the above combination, which is the highest accuracy received so far. And for 'Mountain' class this combination of -features worked tremendously as it gives an accuracy of 97 percent. From those combinations experimented here, this particular shows better accuracy for 'Mountains' compared with others. On combining GLCM with a binary feature BRISK and a feature from MPEG-7 standard, called Scalable colour Descriptor, gives an accuracy of 94.2 percent. But this combination leaves the class, Flower, empty. And it does not give a comparable accuracy for any one of the 10 classes. When applying the GLCM along with SCD, the class, Flower, also get classified, but by the addition of BRISK, it reduces the accuracy. Thus this combination with a descriptor size 492 is discarded. From these experiments one can easily spot out the one combination which shows better overall accuracy, is GLCM + HOG + SCD + FREAK with an accuracy of 96 percent. This produces a descriptor with size 573, as FREAK gives a 64-D feature vector. For the classes 'Africans' and 'Beach', the combination outperforms in terms of precision, recall and accuracy. The system gives a performance, which can stand-by with the existing system as the two of them having the very same accuracy- 96 percent. In our proposed system, the combination of Scalable colour descriptor, colour moment, HOG and GLCM texture features are used. The entire images have undergone training procedure. Annotation was done for around 775 images from 1000.This shows better accuracy for the classes 'Dinosaurs ' and 'Horse' compared with the other experiments. It results no repetition of annotation for the same image.

Table 3. Comparison matrix

Feature Combinations	Overall Accuracy
BRISK + GLCM + SCD	94.20
MSER + HOG + HSV	95.20
HOG + GLCM + HSV	95.50
GLCM + HOG + SCD + FREAK	96.00
GLCM + HOG + SCD + COLOUR MOMENT	96.10

The overall Accuracy related with each of the combination is compared in table 3-'Overall Accuracy'. From the table it is clear that our proposed combination with GLCM, SCD, Colour Moment and HOG gives an accuracy of 96.10 percent. But all the combinations experimented with shows good results in terms of precision, recall and accuracy.The performance of the proposed system was evaluated by the traditional standard methods like precision, recall, F-score and accuracy. The accuracy of 10 classes averaged to compute the overall accuracy of the system. The performance matrix is shown in Table 4 ,

Table 4. Performance matrix

Class Name	TP	FP	FN	TN	Precision	Recall	F-Score	Accuracy
Africans	77	23	23	877	0.77	0.77	0.77	0.95
Beach	76	24	24	876	0.76	0.76	0.76	0.95
Buildings	72	18	28	882	0.80	0.72	0.75	0.95
Bus	90	5	10	895	0.94	0.90	0.91	0.99
Dinosaurs	98	2	2	898	0.98	0.98	0.98	0.99
Elephant	59	20	41	880	0.74	0.59	0.65	0.94
Flower	89	11	11	889	0.89	0.89	0.89	0.98
Horse	83	17	17	883	0.83	0.83	0.83	0.97
Food	72	25	28	875	0.74	0.72	0.73	0.95
Mountain	70	30	30	870	0.70	0.70	0.70	0.94

Table 5. Confusion matrix

AC Vs PC	Africans	Beach	Buildings	Bus	Dinosaurs	Elephant	Flower	Horse	Food	Mountain
Africans	77	1	2	3	0	3	2	5	7	0
Beach	2	76	4	0	0	6	0	1	1	10
Buildings	4	7	72	0	2	6	1	1	3	4
Bus	0	1	3	90	0	1	0	0	5	0
Dinosaurs	0	0	0	0	98	0	0	0	0	0
Elephant	11	4	2	0	0	59	0	4	2	14
Flower	0	0	0	0	0	0	89	5	6	0
Horse	0	4	0	0	0	0	0	83	0	0
Food	6	2	2	2	0	2	8	1	72	2
Mountain	0	5	5	0	0	2	0	0	1	70

The overall effectiveness of the proposed system can be figured out from the table 5 'confusion matrix'. It shows Actual class(AC) Vs Predicted class(PC) comparison. From which it can easily find out how many items are classified correctly and how many are incorrectly. The total number of images which undergone classification can also be counted from the table.

5. CONCLUSION

The proposed system was tried to implement Automatic Image Annotation process by combining features regarding color, texture and a feature for object detection. The fused feature vector with SCD color feature, GLCM texture feature along with HOG and color moments gives a classification result which can stand by with any other existing combinations. The overall accuracy of the system is 96.10%, which is comparable with the other ones. The proposed system is also promising for real time applications because of the precision and recall values. The system can also be enhanced by adding more features or by trial of different other combinations or by another classification algorithms.

6. REFERENCES

[1] A fuzzy K nearest neighbor algorithm, James M KELLER, Michel R,James A givens,IEEE Transactions on systems,man,and cybernetics,vol,SMU-15,NO:4,JULY/AUGUST 1985.

[2] Content Based Image retrieval using color and texture features. International journal on advanced research in electrical, electronics and instrumentation engineering,vol 1,issue 5,November 2015.

[3] Segmentation and histogram generation using the HSV color space for image retrieval.shamik sural.

[4] Evaluating color descriptors for object scene recognition. Transactions on pattern analysis and machine intelligence, vol 10,no:10, July 2010.

[5] Low level feature extraction of an image for CBIR: Techniques and treands,International journal in advanced electronics engineering,vol 1,issue 1.

[6] The MPEG-7 color descriptors, jens-reiner-ohm, leszek, heon jum kin, santhana krishnamachari

[7] Automatic Image Annotation Using Synthesis of Complementary Features, by Sreekumar k, Anjusha B, Rahul Nair, Department of Computer Science College of Engineering Poonjar Kottayam, Kerala, India

[8] Histogram of oriented gradients for object detection, navneet dalal.

[9] Image texture feature extraction using GLCM approach, p.mohanaiah, p.sathyanaranaiah, l.gurukumar, International journal of scientific and research publications, vol.3,issue 5,may 2013.