Classification System for Identifying the Chemical Structure Using Support Vector Machine

P. Santhi[1*], K. Deepa[2]
[1]Associate Professor, Department of CSE, M. Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: santhip.cse@mkce.ac.in
[2]Associate Professor, Department of CSE, M. Kumarasamy College of Engineering, Karur, Tamil Nadu, India.
*Corresponding Author

Abstract: In laboratory, each effort is taken only for identifying the unknown chemicals. All the chemicals are having its own characteristics and structure of molecules such as lines, hexagons and pentagons. The chemical database is used to find the detailed information of that molecule. Even though, the current database does not provide the up to date chemical information. To overcome the above identified problem, this paper introduces the kernel based support vector machine for identifying the chemicals using its structure. The SVM's are becoming more popular algorithm for identification of variety of chemicals in chemical applications. Final result shows the chemical identification and performance analysis of this proposed system.

Keywords: Chemistry, Classification, Molecules, Support vector machine.

I. Introduction

Nowadays, the identification of new chemicals or unknown chemicals are major problems for the chemical engineers [5]. But they all are in position to find the new chemicals to explore the natural world for different applications. The following challenges of chemical engineers' are:

- To develop the effective methods for identifying the dangerous chemicals.
- To identify the name and structure of unknown chemicals.
- To develop and improve the performance of classification system.
- To improve the applications chemistry in all the areas.

To overcome these above challenges, this paper introduces the classification system using support vector machine. This system uses Feature Extraction and Classification for identifying the chemical structure. Feature extraction uses the colour moment for extracting the colour feature and Tamura features are used to extracting the characteristics of chemical image.

During feature extraction, images are analyzed using various colour patterns, texture and shape features. These features are derived from images using feature extraction [4]. Feature extraction is a technique used to find out the attributes as well as characteristics of objects in an image. These characteristics are called as features. These features are used to describe an object. The extracted features must be invariant to the distortions and variations such as translational, rotational and scale of the objects. Some the features are selected using Gabor filter approach [7, 8].

The extracted features are passed into the classifier for identification. Classification plays a very important role in applications such as object classification, military applications, disease diagnosis, etc. It is a process of classifying the real world objects such as buildings, bicycles, faces, etc. Objective of this classification process is to categorise all the pixels in a digital image into one of several classes [6]. The object has been classified or the set of objects have been retrieved based on the salient features of an image such as texture, shape and colour. Image classification is divided into two as follows:

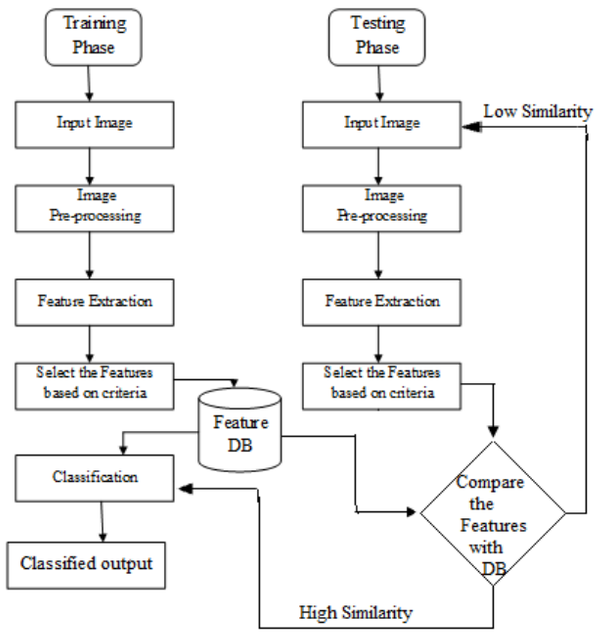

Fig. 1: Steps in Supervised Classification

- Supervised classification.
- Unsupervised classification.

This paper uses the supervised classification for chemical identification. Fig. 1 shows the steps involved in supervised classification.

II. Methodologies

A. Colour Moment for Colour Feature Extraction

Colour moments are the measures that describe the colour distribution of an image and the central moments are used to describe a probability distribution. Main purpose of colour moment is to form the colour indexing for image retrieval applications. The colour moments are invariant for scaling and rotation (Gode *et al*. 2014). It encodes both the shape and colour information. It can be computed for each channel in any colour model. Colour model is a mathematical representation of colour. Here, a colour is represented in the form of a tuple. The tuple contains three or four colour components. The final outcome of this process is colour space. Three colour moments are calculated for each channel. For example, 9 colour moments are calculated for RGB and 12 colour moments for CMYK. Here, the first three order moments are used as features for image retrieval. The first three order moments are as follows:

- The First Order Moment – Mean.
- The Second Order Moment – Standard Deviation.
- The Third Order Moment – Skewness.

i. *Mean*

Mean is the first order colour moment used to find out the average colour in an image. Mathematically it can be represented as follows.

$$E_i = \sum_{j=1}^{N} \frac{1}{N} P_{ij} \tag{1}$$

Where,

N - Number of pixels in an image

P_{ij} - j^{th} pixel of an image at i^{th} colour channel

ii. *Standard Deviation*

Standard Deviation is the second order moment used to calculate the colour distribution in an image. Equation (2) shows the formula to calculate standard deviation.

$$\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij} - E_i\right)^2\right)} \tag{2}$$

Where,

E_i - Mean Value

P_{ij} - j^{th} pixel of an image at i^{th} colour channel

N - Total number of pixels in an image

iii. *Skewness*

It is the third order moment which tells about the shape of colour distribution. It also measures how asymmetric the colour distribution is. Equation (3) shows the formula for calculating the skewness of colour distribution.

$$S_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{N}(P_{ij} - E_i)^3\right)} \tag{3}$$

Where,

E_i - Mean Value

P_{ij} - j^{th} pixel of an image at i^{th} colour channel

N - Total number of pixels in an image

B. Texture Extraction Using Tamura Features with Concentric Circle Based Clustering

This chapter concentrates on tamura features with concentric circle clustering based texture extraction. Here, the images are clustered and tamura features are applied on clustered image to find out the local features of an image. Tamura features are designed based on the human perception of the texture. The tamura features are as follows (Tamura *et al*. 1984):

- Contrast
- Directionality
- Coarseness
- Line Likeness
- Regularity
- Roughness

i. *Coarseness*

Coarseness is used to define the differences between coarse and fine textures. This feature gives the information about size of texture elements. It is one of the fundamental features in textures. It has a direct relationship to scale and repetition rates. It aims to identify the largest size at which the texture exits, even where a micro texture exists (Howarth and Ruger 2004). The coarseness is measured using the following steps:

a) It takes the average at every point over neighbourhoods in the linear size of powers of 2. The average size of the neighbourhood $2^k x 2^k$ at each point (x, y) is represented as follows.

$$Coarseness = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 2^k\, p(i,j) \tag{4}$$

At each point, take the differences between the averages of non-overlapping neighborhoods on opposite sides of a point in horizontal and vertical direction.

The horizontal direction is defined as follows:

$$E_{k,h} = \left| A_k(x + 2^{k-1}, y) - A_k(x - 2^{k-1}, y) \right| \quad (5)$$

The vertical direction is given as follows:

$$E_{k,v} = \left| A_k(x, y + 2^{k-1}) - A_k(x, y - 2^{k-1}) \right| \quad (6)$$

b) Average of these values gives the coarseness of an image.

ii. *Contrast*

Contrast is another important measure in the texture of an image. It aims to capture the dynamic range of gray levels in an image with the polarization of distribution such as black and white. The first step is to measure the standard deviation of an image, and the second step is to find the kurtosis α_4. The contrast measure is defined as follows:

$$Contrast = \frac{\sigma}{(\alpha_4)^n}, n = 1/4 \quad (7)$$

Where,

α_4 is given as follows:

$$\alpha_4 = \frac{\mu_4}{(\alpha)^4} \quad (8)$$

μ_4 - Fourth moment about the mean

σ^2 - Variance

Equations (7) and (8) are used to estimate the contrast feature of an image.

iii. *Directionality*

Directionality is another measure to find out the total degree of directionality. It considers the global property over a region. The angle and magnitude are calculated for each pixel in an image. If the orientation is same, then two images are different. Here, the first step is to calculate the edge direction for each point θ.

$$\theta = \tan^{-1}\frac{\Delta V}{\Delta H} + \frac{\pi}{2} \quad (9)$$

Where,

ΔH, ΔV - Horizontal and vertical derivatives,

It can be calculated as the convolution of input image with 3x3 operators. This operator is represented in Fig. 2.

-1	0	1
-1	0	1
-1	0	1

1	1	1
0	0	0
-1	-1	-1

Fig. 2: 3X3 Convolution Operators

The 'θ' value is used to form a 16 bin histogram at a point. The directionality is calculated using the sum of second order moments of all the histograms.

iv. *Line Likeliness*

It refers to the shape of texture primitives. The texture may have straight line or wave like primitives. The Line - Likeliness is often simultaneous to the direction. Equation (10) gives the computation of line - likeliness.

$$LineLikeliness = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} p_{dir}(i,j) Cos\left[(i-j)\frac{2\pi}{n}\right]}{\sum_{i=1}^{n}\sum_{j=1}^{n} p_{dir}(i,j)} \quad (10)$$

Where,

$p_{dir}(i,j)$ is the 'nxn' local directions of a point d.

v. *Regularity*

It refers to the variations in texture primitives. A regular texture consists of identical textures and it is arranged regularly. An irregular texture is a combination of various texture primitives and it is randomly arranged. Regularity is calculated using equation (11).

$$\text{Regularity} = 1 - n(\sigma_{Coarseness} + \sigma_{Contrast} + \sigma_{directionality} + \sigma_{linelikeliness}) \quad (11)$$

Where,

n – Normalized Factor

σ - Standard Deviation

vi. *Roughness*

It refers to the tactile variations of a physical structure. A rough structure contains angular primitives and a smooth structure contains rounded primitives. The following equation shows the computation of roughness.

$$\text{Roughness} = \text{Coarseness} + \text{Contrast} \quad (12)$$

C. *Fisher Kernel Based Support Vector Machine (FSVM) with Adaboost Classifier*

This method was proposed by Vladimir Vapnik during 1992. It is used for both classification and prediction. This method is more accurate than other classification methods. It is also used in areas such as handwritten image recognition, object recognition and speaker identification.

It is one of the methods for classification of both linear and non linear data. SVM method uses non linear mapping to transform the original data into a higher dimension. In this new dimension, it searches the linear optimal separating hyperplane. This hyperplane is called as a decision boundary for the classes. Support vectors and margins are used to find out the decision boundary for SVM. SVM contains two types of data such as linearly separable data and linearly inseparable data (Qi Gu *et al.* 2009).

The separating hyperplane can be written as follows:

$$W.X + b = 0 \quad (13)$$

Where,

W - Weight Vector

b - Scalar or Bias Value.

For example, X= $\{x_1, x_2\}$, where x1 and x2 are the values of attributes A_1 and A_2.The hyperplane can be rewritten as

$$w_0+w_1x_1+w_2x_2 = 0 \quad (14)$$

Here, w_0 is represented as follows:

$$w_0 = \bar{y} - w_1\bar{x} \quad (15)$$

Where,

$\bar{y}$ and $\bar{x}$ - Mean value of x and y.

$$w_1 = \frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2} \quad (16)$$

The value of w_2 is calculated as shown in equation (17).

$$w_2 = -\left(\frac{w_0 - w_1x_1}{x_2}\right) \quad (17)$$

III. Results of Feature Extraction Using Tamura Features with Concentric Circle Based Clustering

This work is performed using six Tamura features at four directions such as 0^o, 90^o, 180^o and 270^o. These six features are applied on each clustering and the mean is taken for feature selection. The cluster is generated using concentric circle based technique to form the local descriptors of an image using tamura features. Hence, the total number of clusters images is 6 and 5 respectively. The total numbers of features generated from these four directions are 144 and 120 respectively. Therefore, there are 10, 240 features for training set and 5040 for testing set. The values of coarseness, contrast, directionality, line likeliness, regularity and roughness are 33.5865, 0.3181, 0.35835, 0.4298, 0.6891 and 0.53 respectively.

Table I: Feature Extraction

Tamura Features / Directions		Coarseness	Contrast	Directionality	Line Likeliness	Regularity	Roughness
Cluster 1	0^o	31.78	0.078	0.342	0.73	0.024	0.175
	90^o	31.57	0.074	0.307	0.79	0.057	0.163
	180^o	31.83	0.0698	0.321	0.7	0.069	0.184
	270^o	32.01	0.0732	0.316	0.71	0.082	0.172
Cluster 2	0^o	33.94	0.264	0.52	0.623	0.108	0.841
	90^o	32.89	0.274	0.517	0.694	0.157	0.827
	180^o	33.86	0.273	0.537	0.612	0.173	0.819
	270^o	33.52	0.294	0.584	0.639	0.14	0.826
Cluster 3	0^o	34.54	0.416	0.259	0.354	1.654	0.572
	90^o	34.09	0.42	0.237	0.32	1.304	0.543
	180^o	34.95	0.419	0.283	0.371	1.493	0.594
	270^o	34.28	0.438	0.184	0.362	1.285	0.518
Cluster 4	0^o	30.76	0.174	0.138	0.174	0.985	0.739
	90^o	30.89	0.17	0.195	0.157	0.903	0.721
	180^o	30.93	0.173	0.174	0.176	0.924	0.794
	270^o	30.5	0.172	0.425	0.164	0.963	0.729
Cluster 5	0^o	37.56	0.629	0.49	0.298	0.827	0.394
	90^o	37.19	0.618	0.432	0.247	0.884	0.321
	180^o	37.1	0.684	0.483	0.269	0.829	0.35
	270^o	37.54	0.649	0.423	0.206	0.841	0.318

IV. Performance Analysis

Precision and recall are the most familiar measures used to analysis the measurement of ground truth and they are also used to analysis the performance of classifiers. This ground truth is measured by different users with different query images. In each category, there are nine different query images to measure the ground truth of feature extraction and selection.Therefore, totally 27 different types of query images are used to measure the performance of this algorithm. The accuracy is based on the performance of feature extraction and selection.

The precision and recall rate of this proposed technique are 1.5256e-05 and 0.1526 respectively. The accuracy rate of this algorithm is 96.33%. Table II gives the performance, accuracy and execution time of different extraction algorithms.

Table II: Performance of Tamura Vs Concentric Circle Based Tamura Feature Extraction

Tamura Features + SVM	Execution Time	1.90s
	Precision %	91.23%
	Recall %	87.18%
	Accuracy %	90.26%
Tamura Features using Concentric Circle based Clustering (Proposed) + FSVM	Execution Time	1.83s
	Precision %	93.65%
	Recall %	90.72%
	Accuracy %	93.57%

V. Conclusion

Object identification system is an active research area for the past few decades. Object identification is a process of identifying similar objects in an image based on visual features such as colour, texture and shape. This identification system has got huge impact on diagnosis, object detection, education and research. In which, Concentric circle based clustering is used to increase the performance of image clustering. Then, the tamura features are applied on the clustering to capture the local texture information. The Fisher kernel based Support Vector Machine (FSVM) with Adaboost classifier is used for classification which collects the optimized feature subsets for evaluation, fix on kernel functions and it achieves the classification accuracy of 93.6%.

References

[1] Q. Gu, and Z. Song, "Image classification using SVM, KNN and performance comparison with logistic regression," *IEEE Transactions* on *Image Processing,* vol. 13, no. 11, pp. 1179-1187, 2009.

[2] H. Tamura, and N. Yokoya, "Image database systems: A survey," *Pattern Recognition*, vol. 17, no. 1, pp. 29-43, December 1984.

[3] P. Howarth, and S. Ruger, "Evaluation of texture features for content-based image retrieval," In: P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and A. W. M. Smeulders, (ed.) Image and Video Retrieval, CIVR 2004, *Lecture Notes in Computer Science*, vol. 3115, pp. 326-334, Springer, Berlin, Heidelberg, 2004.

[4] O. Ivancicu, "Applications of support vector machines in chemistry," *Reviews in Computational Chemistry*, vol. 23, pp. 291-400, 2007.

[5] J. Park, Y. Choi, A. Min, and W. Huh, "Chemical structure image extraction from scientific literature using Support Vector Machine (SVM)," EECS 545 F07 Project Final Report, 2008.

[6] P. Santhi, and K. Deepa, "Modified boosting classification system for human action classification using 3D modified harris corner detector," *Journal of Advances in Chemistry*, vol. 12, no. 21, pp. 5307-5315, December 2016.

[7] S. Thilagamani, and N. Shanthi, "Object recognition based on image segmentation and clustering," *Journal of Computer Science*, vol. 7, no. 11, pp. 1741-1748, 2011.

[8] S. Thilagamani, and N. Shanthi, "Gaussian and gabor filter approach for object segmentation," *Journal of Computing and Information Science in Engineering*, vol. 14, no. 2, March 2014.