AUTOMATIC MUSIC GENRE CLASSIFICATION USING DUAL TREE COMPLEX WAVELET TRANSFORM AND SUPPORT VECTOR MACHINE

[1]RINI WONGSO, [2]DIAZ D. SANTIKA
[1]Lecturer, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[2]Lecturer, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
E-mail: [1]rwongso@binus.edu, [2]ddsantika@binus.ac.id

ABSTRACT

A fast and accurate automatic method for music genre classification is needed to manage a huge number of digital music currently available. However this is a non-trivial task because musical genres remain poorly defined concepts due to subjective human perception. In this experimental study, an attempt to combine Dual Tree Complex Wavelet Transform (DTCWT) based feature and Support Vector Machine (SVM) as classifier is proposed. The study focused on classifying four genres of music i.e. pop, classical, jazz, and rock by using strong features such as mean, standard deviation, variance, and entropy. Data used in the study is obtained from GTZAN Collection Data Sets. Based on the experimental results, the proposed approach produces 88.33% accuracy, which is higher than several previous methods that could only obtain accuracy up to 86.4%.

Keywords: *Music Genre, Classification, DTCWT, Dual Tree, Wavelet, Support Vector Machine*

1. INTRODUCTION

As digital technology develops, the number of music available in digital format is also increasing. In consequence, to manage this huge number of music, a structured arrangement such as genre is needed. Music genre is a description used to classify music based on characteristics like statistical attributes associated with instruments or rhythms. The huge quantity of music files available is making manual classification not feasible as it is inefficient and very time-consuming. Therefore, a fast and accurate automatic method for music genre classification is important [13]. However this is not an easy task because the knowledge to explain characteristics of music signals is insufficient and too incomplete to allow modeling of computer system that perfectly simulates human music perception [16].

One of the challenges in music genre classification is to find out what it is that allow us to differentiate between music styles which are not directly comparable. Thus, to make comparison possible, data must be first transformed to allow access into essential information contained in them. Result from this process will be features uniquely describing data to be used for classification. This process is referred as feature extraction.

Fourier Transforms (FT) and Wavelet Transforms (WT) are typically used for feature extraction. WT is developed as an alternative of FT which has shortcomings on representing signal with only time axis without the ability to represent frequency and how it varies with time. In general, WT consists of Discrete Wavelet Transform (DWT) and Continuous Wavelet Transform (CWT). DWT has proven to be computationally fast compared to CWT, and in addition, DWT also inherits all the important properties of CWT [9]. Due to limitation of DWT, Dual Tree Complex Wavelet Transform (DTCWT) is developed with the attractive properties of FT, including the nonscilating, nearly shift-invariant magnitude that reduce aliasing, and directional wavelets in higher dimensions with the ability to perform perfect reconstruction (PR) [14], [21] ,[25].

Features obtained from feature extraction will then go through the classification process which can be done using various types of classifier, such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). SVM is superior to ANN and had proven to be adaptive to deal with small number of samples data for training and testing [12]. Other studies also concluded that SVM can give a good classification result [8], [22]. SVM is originally designed for binary classification, and thus can only be used to classify two classes. As solution to this problem, multi class SVM is developed with "one-against-one" being the most superior among other popular methods of SVM [11], [23].

Studies of feature extraction using Wavelets and SVM as classifier have been done since several decades ago with objects such as images. Research for palm print with very successful result had been done with better accuracy when using the combination of dual tree complex wavelets with SVM using Gaussian function kernel [4]. Classification of music using multilinear approaches combined with SVM obtained accuracy of 78.2%, 77.9%, 75.01%, 80.47%, 80.95%, and 78.53%18. Other experiments regarding music genre classification using temporal information and SVM obtained result of 78.6% in classifying 10 music genres [26].

Experiments using DTCWT and SVM had been done with objects of images such as research of recognizing breast cancer using DTCWT and Binary SVM for classification showed 88.64% accuracy as result [27]. By using DWT and DTCWT paired with Neural Network for recognizing breast cancer, high accuracy of 93.33% is produced with data source obtained from Mammographic Image Analysis Society (MIAS) [7]. Recent studies have also concluded that SVM is a leading classifier with the evidence from experiments on classification of e-mail [19] and classification of fruits based on digital photos processed using Multi Class Support Vector Machine with One Against One method and Max-Wins Voting strategy which received accuracy of 88.2% [31].

DTCWT and SVM are so far used for images (2D-data) and show good results in classification. From the experiences mentioned above, it is promising to use combination of DTCWT and SVM for audio signals (1D-data). Moreover, there have been studies of audio classification using image textures [2]. This study aims to propose a music genre classification system with high accuracy using Dual Tree Complex Wavelet Transform and "one-against-one" Multi-Class Support Vector Machine with data sets from GTZAN Genre Collection.

2. DUAL TREE COMPLEX WAVELET TRANSFORM (DTCWT)

As previously mentioned, Wavelet Transform generally consists of Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). DWT is based on sub-band coding and is easy to implement. It also reduces the computation time and resources required. DWT divides signal into two parts, part with low frequency and part with high frequency using two filters which are low pass filter and high pass filter. Some of results from these filters are taken to go through a process called as down-sampling. The whole process is called as decomposition.

The formula for calculating the high pass filter and low pass filter are as follows:

$$y_{low}[k] = \sum_{n} x[n]h[2k-n] \quad (1)$$

$$y_{high}[k] = \sum_{n} x[n]g[2k-n] \quad (2)$$

Result of low pass filter is denoted by $y_{low}[k]$ while result from high pass filter is denoted by $y_{high}[k]$. The initial signal is denoted by $x[n]$, with $g[n]$ as the high pass filter and $h[n]$ as the low pass filter. Decomposition can be done repeatedly until reaching the decomposition desired of level-n. $y_{low}[k]$ is used as the initial signal for every next level of decomposition. This is because according to some researches, result of low pass filter has the important information while the result of high pass filter contains more noise.

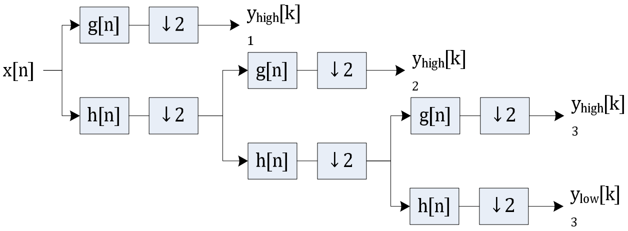

Fig. 1 Three-level wavelet decomposition tree

The original signal of DWT can be reconstructed by concatenating all of the coefficients of $y_{low}[k]$ and $y_{high}[k]$, starting from the last level of decomposition.

DWT has some limitation of such as oscillations of wavelet functions between positive and negative around singularity, shift sensitivity that cause unpredictable change in transforms coefficients, and poor directionality for m-Dimensional Transform, with problem of keeping balance between forward and inverse transforms in reconstructed signal [14], [25].

Dual Tree Complex Wavelet Transform (DTCWT) is introduced with the superiority over DWT in having better directionality and not so influence by shift invariance, and also having perfect reconstruction process. While DWT is employing one tree with original value, the DTCWT is employing two DWTs; meaning that the first DWT is the real part of the transform while the second DWT is the imaginary part.

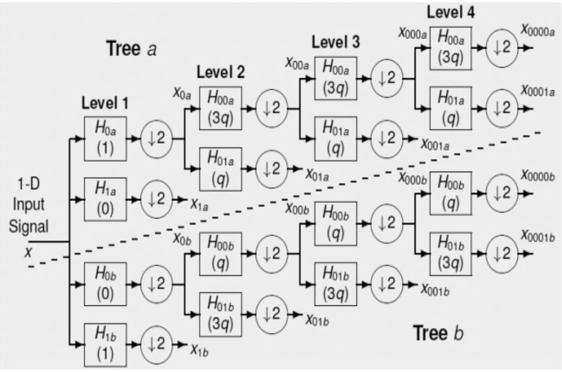

Fig. 2 Dual Tree Complex Wavelet Transform

From the picture above, “Tree a” is real tree and “Tree b” is the imaginary tree which are the two DWTs employed by DTCWT. This is done to overcome the shortcomings of DWT, by doubling the number of samples at each level, but in order to achieve this goal, the samples should come in even number. The filter used in DTWCT is the odd/even filter which has several problems, the two trees have different response frequencies and need biorthogonal filter which trigger the creation of new filter, the Q-shift filter that enables perfect reconstruction. The purpose of Q-shift filter is to get odd/even low pass filter to achieve an extra delay difference of ¼ sample period which satisfies the standard in perfect reconstruction [14].

2.1 Features Analysis

Features extracted from DTCWT must be selected to be used for further process. Several statistical attributes or features commonly used in classifying music genres are: mean [1], [18], [30], variance [1], standard deviation [18], [30], entropy, correlation, contrast, and energy of spectrograms [6], [24] (images) converted from music. Other researches features generated from wavelet coefficients and then use the mean and standard deviation of the features for audio classification [5].

Features commonly used in classifying images [7] are mean, standard deviation, and some of Haralick’s Texture Features that had been optimized [10]. The formula for these features can be seen below.

Mean (3)

$$\mu = \sum_{g=1}^{Ng} g\, P_{(g)}$$

Standard Deviation

$$\sigma = \sqrt{\sum_{g=1}^{Ng} P_{(g)} (g - \mu)^2} \qquad (4)$$

Variance

$$f_4 = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (i - \mu)^2 P_{(i,j)} \qquad (5)$$

Entropy

$$f_9 = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} P_{(i,j)} \log[P_{(i,j)}] \qquad (6)$$

Classification of breast cancer had been done using the texture features of mean, standard deviation, variance, and entropy which are DTCWT features based and defined in the table above. Those features are used because there have been studies that the four statistical attributes are the most significance for images. Using these features, accuracy of 93.3% is obtained [7]. Inspired by the high result, this study will use the four statistical attributes of mean, standard deviation, variance, and entropy.

3. METHODOLOGY

Music genre classification generally consists of two phases of feature extraction and classification. The proposed methodology in this study will be using DTCWT for feature extraction and SVM as the classifier and can be seen in the figure below.

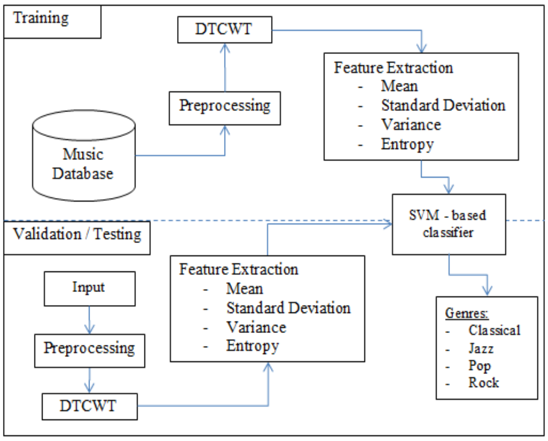

Fig. 3 Proposed Methodology

3.1 Experimental Data

Data used in this study is using the instrumental music gathered in GTZAN Genre Collection Data Sets [28] which consists of 100 tracks for each genre of classic, jazz, pop, rock, blues, reggae, country, hip hop, disco, and metal. Each of the music file in the data sets is 30 seconds long and comes in 22050 Hz Mono 16-bit audio files in .wav format. This data set can be accessed via http://marsyasweb.appspot.com/download/data_sets/.

Based on the taxonomy structures of audio [3] and taxonomy structures [17] of GTZAN Collection Data, four genres of pop, jazz, classical, and rock are chosen to represent music genre classification in this study.

In this methodology, data source will be split into three sets of training, validation, and testing. Training data set will be used as samples to be trained in SVM in order to create optimal model for classification. Validation data set will be used to determine the most optimal constants to do classification while testing data set will be used to evaluate the performance of the proposed methodology.

3.2 Preprocessing and Feature Extraction

Training, validation, and testing data sets will first be prepared for feature extraction through a process called as preprocessing, the process to convert analog music to digital music and sometimes includes the process of recording music using instruments and microphones with the main focus of keeping the naturalness of the sound timbre [15]. Preprocessing will not add any information to the existing data; in fact, it will make extraction of information blocked by noise become possible. Sometimes human voice in music is considered as noise, because it can distract the genre detection process, and due to that, it is much more suggested to just use the instrumental sound. The next part after the music is available in digital format and the noise have been eliminated, is to convert the music into statistical data, usually in format of vector, representing the characteristics of music.

Feature extraction is a method used to get essential information from input data in form of numerical representation which characterize segment of audio. Generally, feature extraction results in big data, and this has caused classification process take very long time. Other than that, computer memory is limited and because of this, it is not possible to process this much data. Therefore, a process called as feature reduction is done to resize data and make it become much smaller to enable processing, but of course, the data used must be reliable, meaning, the result of feature reduction must still be able to represent the real data to create an accurate result. Features obtained from feature extraction, which are mean, standard deviation, variance, and entropy derived from DTCWT result will then be processed for classification using Multi Class SVM with One-Against-One method.

3.3 Evaluation

Evaluation of proposed methodology can be done by doing validation to audio files which have been prepared as validation data. Result of validation will be number of data which is correctly classified and will be displayed using confusion matrix. This number of data will then be computed to accuracy, in percentage, to represent effectiveness of DTCWT.

Table 1. Confusion Matrix for Binary Class [29]

		Predicted	
		Negative	Positive
Actual	Negative	A	B
	Positive	C	D

The elements of the confusion matrix are explained as follow:

- a: the number of correct predictions for the negative samples, called *true negatives*
- b : the number of incorrect predictions for the negative samples, called *false positives*
- c : the number of incorrect predictions for the positive samples, called *false negatives*
- d : the number of correct predictions for the positive samples, called *true positives*

The accuracy is the proportion of the total number of predictions that was correct and it can be determined as the equation below.

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} \qquad (7)$$

4. EXPERIMENTS AND RESULTS

Experiments in this study will be using training, validation, and testing data with 70 files, 15 files and another 15 files respectively for each genres of pop, classical, jazz, and rock. These files are randomly selected once and will be used for all experiments.

Data which have been prepared for training will be converted from audio files (waveform) into numerical values of 1-dimensional vector in pre-processing. The size of the vector is 661,504 x 1, with each value in format of double data type. Preprocessing is also commonly used to remove noise from data before being processed. In terms of audio, background sound, human voice or other sounds that are not the focus of study are considered as noise. Because GTZAN collection data sets come in mono audio, which means this is in form of instrumental music, preprocessing phase to remove noise is skipped as we assume that the data is free from noise.

Feature extraction using DTCWT will convert size of vector with reduction by 2 as follows: in the first level of decomposition, size will be reduced from 661,504 x 1 into 330,752 x 1. After the second level of decomposition, the size becomes 165,376 x 1.

Following level of decomposition, the filter for first level and next level of DTCWT must also be chosen. First level of filter consists of Antonini 9,7 tap filters, LeGall 5,4 tap filters, Near-Symmetric 5,7 tap filters, and Near-Symmetric 13,19 tap filters. Next level of filters are Q-shift 10,10 tap filters, Q-shift 14,14 tap filters, Q-shift 16,16 tap filters, and Q-shift 18,18 tap filters. Experiment has shown that the best result could be obtained by using combination of LeGall 5,4-tap filters and Q-shift 18,18 tap filters (see Fig. 4).

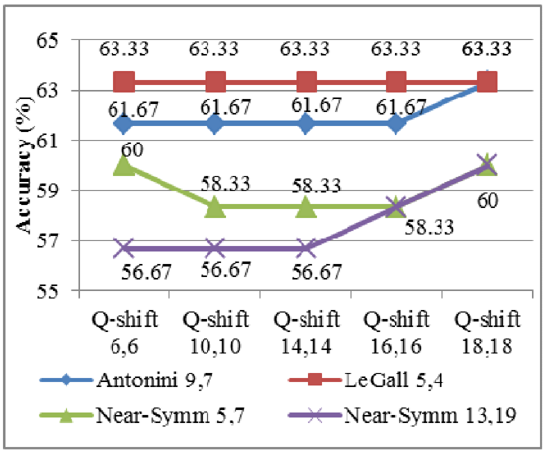

Fig.4 Accuracy Comparison using Different Filters with DTCWT

The next step after feature extraction is to select the potential features by using the statistical attributes value. Accuracy from experiment with only one coefficient and combination of several coefficients of mean, standard deviation, variance, and entropy as statistical attributes derived from DTCWT result can be seen in the figures below.

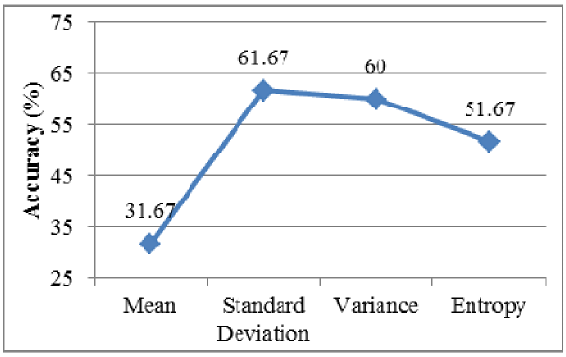

Fig. 5 Accuracy Comparison using only one coefficient with DTCWT

The graph above showed that best result could be obtained using standard deviation. This is because the datas for each class have unique standard deviation value which can be distinguished between each class. Meanwhile, the values of mean for all the classses are similar which lead to incorrect classification.

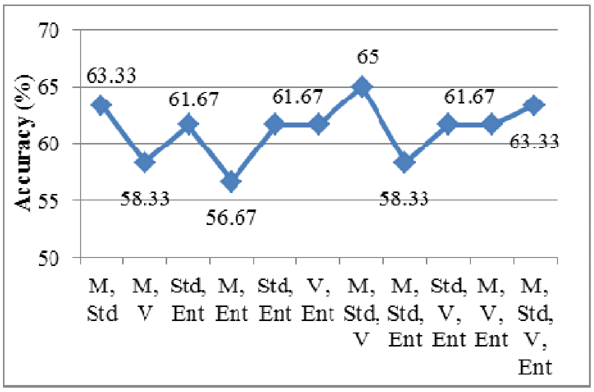

Fig. 6 Accuracy Comparison using combination of several coefficients with DTCWT

The experiments using several combination of mean, standard deviation, variance, and entropy as shown above describe that the most potential combination of attributes to produce high accuracy are by using the combination of mean, standard deviation, and variance which produced accuracy result of 65% followed by the combination of mean and standard deviation and also the combination of mean, standard deviation, variance and entropy (63.33%). Following these results, the combination of mean, standard deviation, and variance are chosen as the attributes that will be used for the experiment in the study. Those attributes are also used in experiment to decide the best configuration for multi class SVM either with one-against-one (OAO) or one-against-all (OAA) method and to define the kernel type and option which are used for supporting the non-linear SVM. The results have shown the same conclusion as previous researches with OAO method to be superior [11], [23]. Meanwhile, the best kernel from the experiment is the Polynomial with option value 2. Results of the experiments are shown in the graph below.

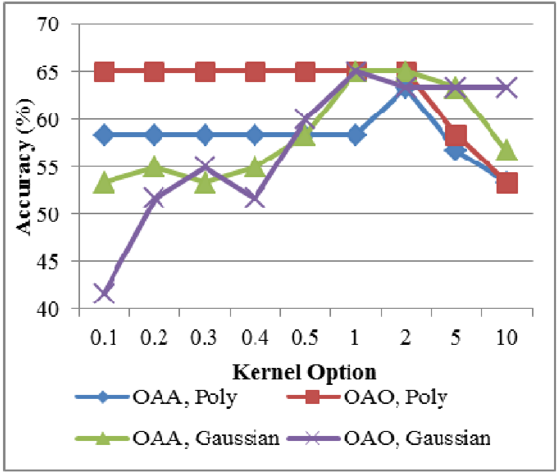

Fig. 7 Accuracy Comparison for SVM Configuration

According to the experiments done, the configuration for DTCWT and SVM for obtaining the best result can be summarized as:

Kernel Option	: Polynomial, 2
Method	: One-Against-One (OAO)
Decomposition Level	: 2
First Level of Filter	: LeGall 5,4 tap filters
Next Level of Filter	: Q-shift 18, 18 tap filters

The experiment with the testing data using these configurations above and additional statistical attributes of mean, standard deviation, and variance has given the result of 88.33% accuracy using DTCWT.

Table 2. Confusion Matrix of Classification based on using DTCWT Features

Genre	Pop	Classic	Jazz	Rock	Total
Pop	**14**	0	0	1	15
Classic	0	**13**	2	0	15
Jazz	0	0	**12**	3	15
Rock	1	0	0	**14**	15

From the confusion matrix shown above it can be seen that DTCWT correctly classified 53 files, 14 of 15 cases of pop, 13 of 15 cases of classical, 12 of 15 cases of jazz, and 14 of 15 cases of rock.

5. CONCLUSION & RECOMMENDATION

Experiments showed that DTCWT-based features with combination of mean, standard deviation, and variance obtained accuracy of 88.33%. This is better than several other methods such as using DWT-based features with the same configuration and treatment as done in the study which generated 83.33% accuracy. The proposed method also has better result compared to past researches which obtained accuracy of 78.2%, 77.9% and 75.01% using different approaches for feature extraction from GTZAN Collection Data Sets combined with SVM as classifier [20]. The proposed method is also superior to research done in classifying genre of music using GTZAN Genre Data Sets with accuracy of 78.6% which also used SVM as classifier [26]. Result in this study is also better than research of using image features which obtained accuracy of 86.4% [2] using the same data sets, GTZAN Genre Data Sets. As shown above, it can be seen that the methodology proposed in this study obtained better result.

Previous research in image classification, as what inspired the method in this study, suggests that no particular type of features is optimal in wide variety of tasks. Experiment in the study showed that by combining multiple features, better results could be obtained. It is also promising, to test the features used in image classification in doing music recognition. Moreover, better result could probably be obtained by combining these features with conventional audio features such as MFCC although it is at the expense of higher computational cost. It should be seen if the results are worth the sacrifice.

Other than the use of features combination, the result will be much more reliable if the research can be done using another data source other than GTZAN Genre Collection Data Set, because this data set were created more than a decade ago. Moreover, the use of just one data set can lead to subjective result. Result from the use of new data set can later be used as accuracy comparison to emphasize the accuracy of this method.

In addition, types of music genre to be classified should also be added in order to keep along with the development of music industry.

REFERENCES:

[1] Anglade, A., Benetos, E., Mauch, M., & Dixon, S. (2010). Improving Music Genre Classification Using Automatically Induced Harmony Rules. Journal of New Music Research 39 (4), 327-339.

[2] Behún, K. (2012). Image features in music style recognition. Proceedings of CESCG 2012: The 16th Central European Seminar on Computer Graphics.

[3] Burred, J. J., & Lerch, A. (2003). A Hierarchical Approach to Automatic Musical Genre Classification. Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03) September 8-11, (pp. DAFX1-DAFXIV). London, UK.

[4] Chen, G. Y., & Xie, W. F. (2007). Pattern recognition with SVM and dual-tree complex wavelets. Journal of Elsevier : Image and Vision Computing Volume 25, Issue 6, 1 June 2007, 960-966.

[5] Chuan, C. H., Vasana, S., & Asaithambi, A. (2012). Using Wavelets and Gaussian Mixture Models for Audio Classification. 2012 IEEE International Symposium, 421-426.

[6] Costa, Y. M., Oliveira, L. S., Koerich, A. L., & Gouyon, F. (2011). Music Genre Recognition Using Spectograms. Systems, Signals and Image Processing (IWSSIP) 18th International Conference on 16-18 June 2011 (pp. 1-4). Sarajevo, Bosnia and Herzegovina: IEEE.

[7] David, & Santika, D. D. (2012). Computer Aided Diagnosis System for Digital Mammogram Based on Neural Network and Dual Tree Complex Wavelet Transform. International Conference on Advances Science and Contemporary Engineering Vol 50, 864-870.

[8] Dhanalakshmi, P., Palanivel, S., & Ramalingam, V. (2009). Classification of audio signals using SVM and RBFNN. Journal Expert Systems with Applications: An International Journal Volume 36 Issue 3, April, 6069-6075.

[9] El-Gebeily, M. A., Rehman, S., Al-Hadhrami, L. M., & Almutawa, J. (2010). Continuous and Discrete Wavelet Transforms Based Analysis of Weather Data of North Western Region of Saudi Arabia. World Journal of Science, Technology and Sustainable Development, Vol 7, No 4, 369-389.

[10] Gipp, M., Marcus, G., Harder, N., Suratanee, A., Rohr, K., Kong, R., et al. (2008). Accelerating the Computation of Haralick's Texture Features using Graphics Processing Units (GPUs). Proceeding of the World Congress on Engineering 2008 Vol 1.

[11] Hsu, C.-W., & Lin, C.-J. (2002). A Comparison of Methods for Multiclass Support Vector Machines. IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 2, 415-425.

[12] Hu, G., & Zhang, G. (2008). Comparison on Neural Networks and Support Vector Machines in Suppliers' Selection. Journal of Systems Engineering and Electronics Vol. 19, No. 2, 316-320.

[13] Jothilakshmi, S., & Kathiresan, N. (2012). Automatic Music Genre Classification for Indian Music. ICSCA Vol 41, 55.

[14] Kingsbury, N. (2001). Complex Wavelets for Shift invariant Analysis and Filtering of Signals. Journal of Applied and Computational Harmonic Analysis Vol 10 No. 3, 234-253.

[15] Kostek, B., & Czyzewski, A. (2001). Representing Musical Instrument Sounds for Their Automatic Classification. Journal of Audio Engineering Society Volume 49 Issue 9, 768-785.

[16] Lee, C.-H., Shih, J.-L., Yu, K.-M., & Lin, H.-S. (2009). Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features. IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 11, NO. 4, 670-682.

[17] Li, T., & Ogihara, M. (2005). Music Genre Classification with Taxonomy. Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05) Volume:5 (pp. 197-200). IEEE International Conference.

[18] Lidy, T., & Rauber, A. (2005). Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification. ISMIR, 34-41.

[19] Lifan, Tao, M., & Hong, X. (2013). The Research on Email Classification based on q-Gaussian Kernel SVM. Journal of Theoretical and Applied Information Technology Vol 48 No 2 ISSN: 1992-8645, 1292-1299.

[20] Panagakis, I., Benetos, E., & Kotropoulos, C. (2008). Music Genre Classification: A Multilinear Approach. International Symposium Music Information Retrieval (ISMIR 2008).

[21] Priya, R. A., Narote, S. P., & Patil, A. V. (2011). Dual Tree Wavelet Transforms in Image Compression. International Journal of Engineering Sciences Research (IJESR) 1 (1).

[22] Rao, C. S., Kumar, S. S., & Mohan, B. C. (2010). Content Based Image Retrieval Using Exact Legendre Moments and Support Vector Machine. The International Journal of Multimedia & Its Applications (IJMA) Vol 2 No 2, May, 69-79.

[23] Reddy, E. K., & Bellary, J. (2012). Multi-Class Support Vector Machines - A Comparative Approach. International Journal of Applied Physics and Mathematics, Vol. 2, No. 4, 264-265.

[24] Scaringella, N., & Zoia, G. (2005). On the modeling of time information for automatic genre recognition systems in audio signals. Proceedings of the 6th Int. Symposium on Music Information Retrieval, London, UK.

[25] Selesnick, I. W., Baraniuk, R. G., & Kingsbury, N. G. (2005). The Dual-Tree Complex Wavelet Transform. IEEE SIGNAL PROCESSING MAGAZINE, 123-151.

[26] Tao, R., Li, Z., & Ji, Y. (2010). Music genre classification using temporal information and support vector machine. ASCI Conference Vol. 77, 78.

[27] Tirtajaya, A., & Santika, D. D. (2010). Classification of Microcalcification Using Dual-Tree Complex Wavelet Transform and Support Vector Machine. 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies.

[28] Tzanetakis, G., Essl, G., & Cook, P. (2001). Automatic Musical Genre Classification of Audio Signals. ISMIR - The International Society for Music Information Retrieval.

[29] Vercellis, C. (2009). Business Intelligence: Data Mining and Optimization for Decision Making. Milan, Italy: John Wiley & Sons.

[30] Wagner, J., Kim, J., & Andr´e, E. (2005). Implementing and Comparing Selected Methods for Feature Extraction and Classification. Multimedia and Expo: ICME 2005 on IEEE International Conference, 940-943.

[31] Zhang, Y., & Wu, L. (2012). Classification of Fruits Using Computer Vision and a Multiclass Support Vector Machine. Sensors 2012 Vol 12 ISSN 1424-8220, 12489-12505.