Gamelan instrument sound recognition using spectral and facial features of the first harmonic frequency

Aris Tjahyanto\textsuperscript{1,*}, Diah Puspito Wulandaran\textsuperscript{2,+}, Yoyon K. Suprapto\textsuperscript{2,\textsuperscript{y}} and Mauridhi Hery Purnomo\textsuperscript{2,\textsuperscript{x}}

\textsuperscript{1}Information Systems Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
\textsuperscript{2}Electrical Engineering Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

(Received 7 August 2013, Accepted for publication 19 August 2014)

Abstract: Principal component and spectral-based feature sets were applied to the recognition of gamelan instrument sounds using support vector machines (SVMs). The principal components were calculated on the basis of a segmented scalogram from the first harmonic frequency of the gamelan recordings. The segmented scalogram is assumed as a “facial image” of the gamelan instrument sound in a frontal pose, neutral expression, and normal lighting. The scalogram was computed from the gamelan sound signal using a continuous wavelet transform (CWT). The performance and contribution of the principal component and spectral-based features were compared using an F-measure. For the training phase, the feature sets were extracted from isolated tones that were recorded over the entire frequency range of four gamelan instruments (demung, saron, peking, and bonang families). Using 90%/10% splits between the training and validating data sets, model classifiers were constructed from the radial basis function (RBF) kernel SVM. The classifiers are composed of 28 separate One-Against-One multiclass classifiers. The experiment showed that the spectral-based feature set shows an average F-measure of 74.05% and the appearance-based feature yields 71.87%. For \textit{saron}-only note tracking, the spectral-based feature set had an F-measure of 83.79%, higher than the \textit{demung}-only note tracking, which yielded 63.89%.

Keywords: Support vector machines, Automatic transcription, Wavelet transform

PACS number: 43.60.Lq [doi:10.1250/ast.36.12]

1. INTRODUCTION

The process of converting a gamelan audio signal into \textit{balungan gendhing} notation can be done automatically by recognizing which \textit{balungan} instrument is playing which note. \textit{Balungan} is the skeleton or core melody of a \textit{gendhing}, which appears like the one shown in Fig. 1; \textit{gendhing} is a generic term for any gamelan composition. The gamelan instruments that are played imitating the core melody are known as \textit{balungan} instruments, such as \textit{demung}, \textit{saron}, and \textit{peking}. The \textit{bonang} plays as a leading and elaborating instrument. The \textit{demung}, \textit{saron}, and \textit{peking} play \textit{balungan} within their one-octave ranges. The \textit{demung}’s octave overlaps with the lowest octave of \textit{bonang}, and the \textit{saron}’s octave overlaps with the highest octave.

Gamelan has distinctive scale systems and intervals. Gamelan also has no standard tuning system. Gamelan tuning practice was not governed by concern with the mathematical purity of interval vibration ratios [1]. For example, a gamelan set from Yogyakarta differs from a gamelan set from Surakarta. Because of the distinctive scale systems and their intervals, the frequency range and the intervals may vary slightly.

Each frequency channel that represents a note contains many onsets from several instruments played in a gamelan ensemble. Therefore, the transcription processes for a multi-instrumental gamelan that produces \textit{balungan gendhing} needs information about what instruments attributed to the onset. To differentiate one gamelan instrument from another, we used features extracted from the time-frequency domain. After segmenting the signals in the time-
frequency domain, we can recognize the gamelan instrument using the global representation of the segment or distinctive spectral features such as centroid, flux, roll-off, skewness, and kurtosis.

In the facial recognition field, it is common to implement principal component analysis (PCA) or linear discrimination analysis (LDA) for appearance-based or global representations of the face image. It is also common to use geometric relationships among the facial features and to use facial features such as eyebrows, eyes, nose, mouth, and cheeks. In the field of automatic music transcription systems, monophonic transcription tasks rely on information in the time domain. Polyphonic transcription tasks, by contrast, depend on the analysis of information in the frequency domain. In the time domain or frequency domain, many acoustic features are captured for instrument recognition.

In this paper, we propose a novel feature extraction approach based on PCA and spectral-based features to convert the “facial image” of a segmented scalogram taken from gamelan instrument tones. Then we apply support vector machines (SVMs) to build a classifier and evaluate the performance of those features in gamelan instrument recognition.

1.1. Previous Work

PCA has been known as an effective technique for feature extraction in the field of facial recognition [2]. Many researchers have also addressed the use of LDA as a feature in facial recognition research [3]. PCA also has been applied to image compression [4] and object detection [5]. De Paula showed that PCA can be used to represent the different timbre spaces of a musical instrument [6]. Kitahara implemented PCA for reducing a 129-dimensional feature space to a 79-dimensional one, then applied LDA to further reduce the feature space to a lower-dimensional one [7].

In automatic music transcription, many approaches have been applied to estimate the pitch and instruments. Hidden periodicities in a time domain signal can be determined using an autocorrelation function. The peaks are related to the lags where periodicity is stronger [8]. Suprapto et al. introduced a technique for generating gamelan transcription using a spectral density model and adaptive cross correlation [9]. These approaches used the energy profile taken from the frequency channel of a saron instrument. These approaches work well for producing gamelan notation taken from saron-only notes. One of the disadvantages of these approaches is the difficulty in distinguishing the energy profile of the saron and bonang signals because of the overlapping frequency channel of both instruments.

Another approach is the pattern recognition technique. This technique requires that a set of features be extracted from the audio signal [10,11]. The feature set for the recognition process can be grouped into spectral-based features and automatic speech recognition (ASR) features [12,13]. The common features for audio signal are zero crossing rate, envelope, RMS energy, centroid, spectrum representation, and flux [14,15].

Feature extraction and selection is a key process in automatic music transcription using pattern recognition technology. Feature extraction is a process for discovering a set of vectors that represents an observation while reducing the dimensionality. Many algorithms have been developed to transform an audio signal into another representation for extracting the feature. Common methods for transforming the audio signal are fast Fourier transform (FFT), short time Fourier transform (STFT), discrete wavelet transform (DWT), and continuous wavelet transform (CWT). On the basis of the transformed audio signal, various features are calculated and extracted.

1.2. Proposed Method

The goal of this paper is to compare the performance of principal component and spectral-based features for gamelan instrument sound recognition, especially the instruments that imitate balungan gendhing. The principal component obtains an orthogonal projection of the data signal, including the noise component. The spectral-based features take into account only relevant features if the process of feature selection has been done carefully. The performance of the principal component and spectral-based features is cross-validated using SVMs. There are two main reasons for addressing these tasks using SVMs. First, accurate recognition of gamelan instruments is itself important for automatic transcription. Second, because of the effectiveness of SVMs [16] they recently became one.
of the most popular recognition and classification methods. They have been used in a wide variety of applications, such as text classification [17], facial recognition [18], and gene analysis [19].

The rest of this paper is organized as follows. Section 2 describes the system architecture of automatic gamelan notes or balungan gendhing using SVMs as classifiers. Section 3 presents our experiments and discusses the results. Finally, Sect. 4 discusses the conclusions of our experiments.

2. SYSTEM ARCHITECTURE

The process of automatic gamelan transcription aims to convert a gamelan audio signal into gamelan notes. Gamelan notes are any system that represents the pitch of a gamelan sound, through the use of written symbols. To produce precisely the balungan gendhing notations from a gamelan sound signal, it is necessary to have information such as the onset, the pitch, and which instrument produced the tone.

Figure 2 shows the architecture of the Automatic Gamelan Notes Transcription system. The Pre-Processing module performs the cleaning of the gamelan sound signals (mono or stereo). The WAV file is then processed by the Time-Frequency Signal Representation module for producing the spectrogram. The Frequency Channel Estimation module determines the significant frequency channel of the signals from the individual characteristics of the gamelan ensemble. The determined significant frequency channel is used by the Segmentation module for performing onset identification and segmentation in the time-frequency domain and then continued for feature extraction. Based on the extracted features, the Gamelan Instrument Classifier module estimates the candidate pitch values and the predicted instruments. The Cleaning and Tabulation module deletes uncommon events and then tabulates the estimated notes into a balungan gendhing notation.

2.1. Time-Frequency Signal Representation Module

For determining the significant frequency channel and the onset, we used STFT because of its efficient computational cost [20]. However, STFT suffers from a common shortcoming that the window length influences the resolution of the spectrograms [21,22]. For a pitch analysis, such as automatic gamelan note transcription, the frequency resolution of the spectrogram is more important than the time resolution [22]. Thus, STFT with a long window is a powerful tool for automatic gamelan note transcription. From a gamelan sound signal, the STFT is computed. With a sampling rate 44.1 kHz, a Hamming window with 4096 length and an overlap of 60% was used to ensure a perfect spectrogram for determining the significant frequency channel. The result of STFT can also be plotted on a 3D surface (as shown in Fig. 3) as a function of time and frequency. The magnitude is represented as the height of a 3D surface spectrogram.

We used CWT for calculating the scalogram of the gamelan sound signal for feature extraction because of its flexibility in time-frequency resolutions. The flexibility of the time-frequency resolution is based on the window or mother wavelet function. For small scales or high frequencies, the wavelet transform produces good time resolution and poor frequency resolution. For large scales or low frequencies, it gives good frequency resolution and poor time resolution.
For performing gamelan instrument recognition, the mother wavelet should have a good frequency resolution in order to differentiate the significant frequency channel with high accuracy. It is better if the computing process for producing the scalogram can be done in a short time. We chose Complex Morlet as a mother wavelet in the analysis of the gamelan sound signal because its characteristics fulfill the requirements [23].

2.2. Frequency Channel Estimation Module

Gamelan instruments are constructed by hand. They are tuned according to the constructor’s sense and experiences to definite pitches corresponding to the pelog or slendro scale. The sizes of the intervals of adjacent tones are not standardized, and the tuning of the intervals varies from one gamelan set to another. This leads to individual characteristics or variations of fundamental frequency in gamelan sound signals [24].

Because of the individual characteristics of a gamelan, the significant frequency channels should be determined from a gamelan recording. We determined the significance level of a frequency channel using the sum-of-energy profile intensity. The sum-of-energy profile intensity for a frequency channel can be defined as the area under an envelope (Aue). When the Aue is high, the frequency channel is important, and the significant frequency channel can be determined using Alg. 1.

Using the spectrogram, we investigated the energy profile and calculated the area under the envelope $Aue(j)$ for each frequency bins where $j = 1, 2, \ldots, Q$ is the frequency bin index. After we determined the indexes of the significant frequency channels using Alg. 1, all the indexes were sorted in descending $Aue(j)$ order. The higher the energy profile is, the more important the frequency channel. The $N$ first of the sorted $Aue(j)$ are assumed as candidate frequency channels of the gamelan sound signals. The $Aue(j)$ for each frequency bin can be calculated using Eq. (1), where the $A_{ij}$ is amplitude of the time domain envelope in frequency bin $j$, where $i = 1, 2, \ldots, P$ is the index of a signal with a length of $P$.

$$Aue(j) = \sum_{i=1}^{P} A_{ij}, i = 1, 2, \ldots, P \quad (1)$$

After selecting a frequency index, Alg. 1 sets the $Aue(j)$ to 0 for $j = j_{\text{max}} - r, \ldots, j_{\text{max}} + r$. The value of $r$ can be defined as range of index frequencies for the significant frequency channel. The value of $r$ can be estimated using Eq. (2).

$$r = j_{t} - j_{\text{LBL}} = \left\lfloor \frac{W \times C_{1} \frac{f}{f_{s}}}{} \right\rfloor \quad (2)$$

where $j_{t}$ and $j_{\text{LBL}}$ are the index frequency for the center frequency and the lower band limit frequency respectively, $W$ is the length of the Hamming window, $C_{1}$ is a constant number, $f = j \times f_{s}/W$ is the frequency bin from STFT and $f_{s}$ is the sampling rate of the gamelan sound signals. If $j_{t} = Wf_{s}/f_{s}$ and $j_{\text{LBL}} = W(1 - C_{1})f_{s}/f_{s}$ then we obtain Eq. (2). The value of $C_{1}$ is determined such that the frequency channels do not overlap each other. Using $C_{1} = 0.06$ is sufficient to cover each of the frequency channels from the lowest key of demung to the highest key of peking (see Fig. 4).

The value of $N$ is determined using Alg. 2. It depends on the song and the gamelan instruments played. The selected frequency channel is then validated; if a frequency channel is in the range of the gamelan frequency, then the channel will be involved in the next process.

---

**Algorithm 1: Determining the significant frequency channels**

**Data:** Spectrogram $M(j,i)$ from the signals, $N$ maximum number of the frequency channel should be retrieved and vector frequency $F(j)$; where $i = 1, 2, \ldots, P$ is the index of signal with a length of $P$ and $j = 1, 2, \ldots, Q$ is the frequency bin index.

**Result:** Significant freq. channels index SFCidx.

1. Set $jCnt \leftarrow 0$, SFCidx $\leftarrow \emptyset$;
2. Calculate $Aue(j)$ for $j = 1, 2, \ldots, Q$ using Eq. 1;
3. repeat
4. Calculate $AueSUM = \sum_{j=1}^{Q} Aue(j)$;
5. Get $[jmax, jmax] \leftarrow \max(Aue(j))$, for $j=1,2,\ldots, Q$;
6. Get $f \leftarrow F(jmax)$;
7. Get the value of $r$ of frequency $f$ using Eq. 2;
8. Set $M(j,i) \leftarrow 0$, $Aue(j) \leftarrow 0$, for $i = 1, 2, \ldots, N$ and $j = jmax - r, \ldots, jmax + r$;
9. Set $jCnt \leftarrow jCnt + 1$;
10. Set SFCidx $\leftarrow$ SFCidx $\cup$ jmax;
11. if $\{AueSUM = 0\}$ break;
12. until $jCnt \geq N$;
2.3. Segmentation Module

The first step in obtaining the onset for a recorded gamelan sound signal is to determine the significant frequency channel. Then we performed pre-processing to enhance the energy profile or envelope for each frequency channel. The enhancement processes consist of smoothing and thresholding. Those processes are needed for increasing the onset identification performance.

An onset can be defined as the instant when the attack transient begins or when a new note begins [21]. The note onset can be identified by sudden changes in the energy profile [25]. In the case of gamelan note attacks, this increasing energy will be very sharp. Taking the first order of the energy profile, the peaks will show the onset locations [26]. Sometimes there is noise caused by a double candidate onset at a close distance. This shortcoming can be eliminated by using a minimum distance between onsets. Practically, the minimal distance between onsets is around 250–500 ms, depending on the song type or genre. Because the fastest gamelan tempo is 250 ms [9], we can remove one of the onsets from the list if the distance between onsets is less than 250 ms.

Figure 5 shows the segmented time-frequency representation from the piece Manyar Sewu that was played using saron and bonang. Each segment is part of a spectrogram corresponding to the individual gamelan event. Each segment of the spectrogram represents one “facial image” of gamelan events. The segmentation was performed using the onset information and note duration. The segment captures only the “facial image” of the first harmonic; the detail of the segment can be seen in Fig. 6. It shows the contour plot and scalogram of one segment taken from the bonang pitch 6 low, which has a center frequency of 463 Hz.

To speed up the CWT for producing the scalogram, we selected as few the scales as possible. If a frequency $f$ is one of the significant frequency channels for a gamelan recording, the scales for wavelet wave and frequency $f$ can be estimated using Alg. 3. This technique ensures that only the relevant time-frequency representation will be produced by the CWT.
The Feature Extraction module performs an extraction on the segmented scalogram. The segmented scalogram is a matrix with size $n$-by-$m$. The $n$ rows correspond to frequency bins, and the $m$ columns to the duration or observation. It is common that bonang has a short duration or sustain and demung has the longest duration compared to the other instruments (saron and peking). We set the duration at 630 ms to capture the "facial image" of the gamelan instrument sound in a frontal pose, neutral expression and normal lighting. The results were then rescaled into a value range from $-1$ to 1. Appearance-based features were compared to the spectral-based features set, which has been previously developed for gamelan instrument recognition [27]. Using those features, the SVMs were used in the training and testing phase. This study compares the recognition performance of SVMs with the optimal kernel parameters determined using a grid-search algorithm [27,28].

### 2.4.1. Appearance-based features

The appearance-based approaches use PCA to extract the holistic appearance features of a "facial image." PCA is a technique for extracting relevant information and reducing the dimensionality of the data sets. The technique generates a set of variables, known as principal components (PCs). All the PCs are orthogonal or uncorrelated to each other. The PCs can be determined by calculating the eigenvectors and eigenvalues of the data covariance matrix. For the principal components to work properly, the data should have a zero-mean [29].

When onset detection has been done, the scalogram is fragmented into smaller sections or segments and a feature set is extracted. The scalogram was segmented into small blocks, producing a matrix $(n \times m)$ for each segment. Using the data matrix, the appearance-based features of each segment were extracted and calculated. The algorithm for computing the appearance-based features of a data matrix $X$ can be seen in Alg. 4.

### 2.4.2. Spectral-based features

Spectral-based features are various features that are extracted from the time-frequency representation of a signal to describe its spectral nature, such as centroid, flux, roll-off, skewness, kurtosis, and other statistical descriptions of the magnitude spectrum. We extracted 34 spectral features, as shown in Table 1.

Spectral centroid is a measure of the center of gravity of a spectrum. It is calculated by multiplying the value of each frequency by its magnitude in the spectrum, then...
taking the sum of all these. The value is then normalized by dividing it by the sum of all the magnitudes. Spectral flux is an indication of the degree of change of the spectrum and defined as the squared difference between the normalized magnitudes of successive spectral distributions. Spectral roll-off is a measure of spectral shape, and defined as the frequency below which 85% of the magnitude distribution of the spectrum is concentrated [30]. Spectral kurtosis $K$ is a measure of the peakedness or flatness of the shape of the power spectrum distribution. A positive kurtosis $K > 3$ indicates a peaked distribution, the standard normal distribution has a kurtosis $K = 3$, and negative kurtosis $K < 3$ indicates a flatter distribution [31]. In this work, we used spectral roll-offs 85% and 40%. The features were calculated and additional features were appended including statistical properties such as skewness, mean, and variance of the spectrum.

### 2.4.3. Appearance- and spectral-based feature combination

It is quite possible that better performance will be gained by using a feature combination from several feature sets. The feature combination for appearance- and spectral-based features can be arranged into two subsets. The first subset is APP&SPC, a direct combination from the appearance- and spectral-based features. The second subset is SELECTED, the combination of the top $n$ features from the appearance- and spectral-based features. The ranking process was performed using the Gain Ratio technique provided by the Weka data mining tool [32].

### 2.5. Gamelan Instrument Classifier Module

The note for a scalogram segment can be determined from the fundamental frequency that is indicated by the highest energy intensity. Using the identified fundamental frequency, we obtain information about the note and select the proper classifier for predicting the gamelan instrument. The classifier decides whether the segment belongs to the demung, saron, peking or bonang class on the basis of the information learned during the training phase.

Our Gamelan Instrument Classifier is composed of 28 separate One-Against-One (OAO) multiclass classifiers, based on an SVM with an RBF kernel (as shown in Fig. 7). Each individual classifier is intended for one note or frequency channel, starting from the lowest note of demung up to the highest note of peking in the slendro scale. During the training phase, each classifier was trained using audio features extracted from a segmented scalogram with the same note or fundamental frequency.

Supervised training of a classifier, which utilizes audio features to predict the gamelan instrument, requires a labeled set of audio features. The training data are composed of audio features extracted from each isolated tone for all four gamelan instruments, starting from the lowest note to the highest note. In the training phase, some support vectors are selected to separate the gamelan instrument classes with the largest possible margin and fewest misclassifications by the appropriate hyper parameters.

Before performing the training phase, we determined the appropriate hyper parameters for each classifier. For the RBF kernel, we should determine the best value for each hyper parameter pair $(C, \gamma)$. For choosing the best value, we perform a grid-search using the set of values $\log_2(C) \in \{-4, -2, \ldots, 10\}$ and $\log_2(\gamma) \in \{-6, -5, \ldots, 14\}$ [28].

### 2.6. Cleaning and Tabulation Module

The Cleaning and Tabulation module produces the balungan gendhing which is the transcription of the audio source: a list of pitches, onset times, and identified instruments of all detected notes. Before producing balungan gendhing, the module performs a cleaning process to delete unwanted events. The unwanted events can be double onsets at a close distance, or a false event caused by a high-intensity overtone.

---

**Table 1** Spectral-based features.

| No | Features                        | Number |
|----|---------------------------------|--------|
| 1  | Fundamental frequency           | 1      |
| 2  | Spectral centroid               | 1      |
| 3–4| Spectral roll-off               | 2      |
| 5  | Spectral flux                   | 1      |
| 6  | Spectral skewness               | 1      |
| 7  | Spectral moment                 | 1      |
| 8  | Spectral kurtosis               | 1      |
| 9  | Spectral entropy                | 1      |
| 10 | Spectral slope                  | 1      |
| 11 | Spectral bandwidth              | 1      |
| 12 | Mean                            | 1      |
| 13 | Standard deviation              | 1      |
| 14 | Mode                            | 1      |
| 15 | Median                          | 1      |
| 16 | Variance                        | 1      |
| 17–25| Percentile                      | 9      |
| 26–34| Quantile                        | 9      |

---

**Fig. 7** A Gamelan Instrument Classifier Model. A classifier is intended for a note, starting from the lowest pitch of demung up to the highest pitch of peking in the slendro scale.
Here, tabulating means the process of arranging the identified pitches or tones in the gamelan cipher notation or balungan gendhing. The balungan gendhing is written using numbers to represent the notes, which are arranged in groups called gatra. A gatra is the shortest unit and consists of four beats, which may be occupied by either a note number or a rest (represented as a dot). The balungan notation can be reconstructed from the gamelan sound recording by choosing the notes played using the saron family.

3. DATA SETS AND EXPERIMENTAL RESULTS

In this section, the experimental tests that were set up to assess the performance of the proposed method are described. First, the data sets that were used in training and testing are shown. Then, the evaluation parameters are defined. Finally, the results of the comparison of the proposed method are presented.

3.1. Data Sets

For the experiments, we used recordings of four Javanese gamelan instruments (see Fig. 8); demung, saron, peking, and bonang, to generate training, testing, and validation data sets. We used isolated tones for data training and validation that were performed over the entire pitch ranges of the instruments. A list of recording titles for the testing phase can be seen in Table 2.

The peking, which has narrow thick bars, presents the highest octave [9,33].

The bonang is composed of two rows of small gongs, placed horizontally on cords stretched over the rectangular wooden frame. The bonang is played with two padded sticks. The bonang family can be categorized into two groups, the bonang barung and bonang panerus. The bonang barung has the same octave as the demung and saron, whereas the pitch of bonang panerus occupies the same octave as the saron and peking.

The gamelan sounds were recorded (monaural) at a sampling rate of 44.1 kHz. We produced the sound data by randomly hitting the keys or bars of metal at the center, upper, and lower areas [34]. For producing the sound samples, each bar or key was struck with its own hammer around 80 to 100 times. In total we obtained about 4894 isolated tone samples, with various intensity levels and across the entire pitch range of the four instruments.

We randomly partitioned the data sample into training and validation data sets to verify the performance and robustness of the classifiers. The validation data set is completely different from the training data set. Most of the data were used for training, and a small portion of the data were used for validation. The number of instances for each class is approximately equal, depending on the generation of index vectors for cross-validation of the instances by randomly selecting a portion of the data. Each class consists of one tone from the entire pitch range of the four instruments; this amounted to 33 classes.

The gamelan recordings were selected with the purpose of creating a heterogeneous data set, including demung-saron, saron, saron-peking, saron-bonang, demung-saron-peking, demung-saron-peking-bonang, and other combinations.

| No | Title        | Length (sec) | Instruments          |
|----|--------------|--------------|----------------------|
| 1  | Manyar Sewu  | 29.18        | S                    |
| 2  | Manyar Sewu  | 76.80        | S, P                 |
| 3  | Manyar Sewu  | 74.66        | D, S                 |
| 4  | Manyar Sewu  | 41.62        | S, P                 |
| 5  | Manyar Sewu  | 121.20       | D, S, P, BG          |
| 6  | Manyar Sewu  | 37.66        | S, BG                |
| 7  | Singo Nebah  | 92.21        | D, S, P, BG          |
| 8  | Manyar Sewu  | 47.70        | D, S, BG             |
| 9  | Manyar Sewu  | 46.18        | D, S, BG             |
| 10 | Manyar Sewu  | 48.88        | D, S, BS             |
| 11 | Bengawan Solo| 133.80       | D, S, P, BG          |
| 12 | Caping Gunung| 157.30       | D, S, P, BG          |
| 13 | Manyar Sewu  | 47.70        | D, BG                |
| 14 | Manyar Sewu  | 48.88        | D, BS                |
| 15 | Manyar Sewu  | 50.54        | P, BG                |
3.2. Evaluation Parameters

We performed a statistical evaluation of the performance of the gamelan instrument sound recognition method. The results are summarized by three statistics: the Precision, the Recall and the F-measure. Calculating these statistics requires the TP, FP, and FN parameters. TP is the number of the true positives, where the instrument was correctly recognized. FP is the number of false positives, where another instrument was recognized as the actual instrument. FN is the number of false negatives, where an instrument was recognized as another instrument.

- **Precision** is the ratio of correctly recognized pitches and instrument to all recognized pitches and instrument. Precision represents the percentage of correct positive recognitions.

\[
Precision = \frac{TP}{TP + FP}
\]  

- **Recall** represents the capacity of the proposed system to recognize the positive examples.

\[
Recall = \frac{TP}{TP + FN}
\]

- **F-measure** is the harmonic mean of Precision and Recall. It conveys information about the balance between FP and FN.

\[
F-measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}
\]

3.3. Discussion

The proposed method presents a novel feature that is useful for recognizing gamelan instruments that have variations because of its handmade construction and different tuning systems. The proposed method is a holistic appearance-based feature extraction that takes advantage of the “facial image” of a gamelan instrument’s sound signals. This approach represents global information about the “facial image” and provides practical advantages, such as the ease of explanation and interpretation of the semantic meaning of the “facial image” for given signals.

Figure 9 illustrates the “facial image” of a segmented scalogram of a gamelan sound signal. The segmentation process was performed on the basis of the onset information. For experiments that deal with scalogram segmentation, the results show segmentation instability. The instability of segmentation causes the “facial image” to be shifted horizontally along the x-axis or vertically along the y-axis. In this experiment we ignore the shifting effect and use it to explore the best extraction techniques that are suitable for gamelan instrument sound recognition.

The shifting along the x-axis was caused by the smoothing procedure that was applied before the peak-picking algorithm in the onset identification process. The smoothing process is intended for removing spikes and ripples noise. We used a generalized moving-average Savitzky-Golay filter to remove the noise, with a span of about 1% of the total number of data points. However, the smoothing process has side effects, and under certain condition causes the identified onset time to be earlier than the actual onset time.

The “facial image” shifting also takes place along the frequency or y-axis. This shifting is caused by a poor procedure for significant frequency channel identification. However, the shifting along the y-axis tends to cause consistent results for the same frequency channel. Thus the consequence of the shifting along the frequency axis in the feature extraction result is still predictable, after extracting the energy profile and calculating the energy’s first order derivative. The next steps for identifying the onset time are the smoothing, thresholding, and peak-picking techniques applied to the first derivative of the energy profile. Figure 10 shows the identified onset as a gray color in a piano-roll style with the vertical axis representing frequency, ranging from 0 Hz (at the bottom) to 1,200 Hz (at the top). The horizontal axis is time (in seconds).

Every identified onset is matched with every corresponding onset of the ground truth reference as shown in Fig. 10. The ground truth is created semi-automatically using the following steps: (i) significant frequency channel identification, (ii) onset identification and verification, and (iii) manual annotation by gamelan experts. To capture a “facial image” of one segment, we set the duration for each event at around 630 ms, corresponding to the saron average duration. The bonang has a very fast decay
behavior, with an average duration around 240 ms. The average duration is 570 ms for the peking and 3790 ms for demung. However, when playing saron and demung, the player usually applies a damping technique to stop the sound by holding the key. This technique reduces the note duration. The average duration of saron is too long to capture the “facial image” of bonang, but long enough to capture that of demung. The results show that the onset identification algorithms exceed the ground truth reference.

The performances of the appearance-based (APP) and spectral-based (SPC) features for gamelan instrument sound recognition are compared. The comparison is made on the basis of a gatra, the shortest unit of gamelan compositions that consists of four basic beats [33]. Every gatra of the identified onset is matched with every corresponding gatra of the ground truth reference of each audio sample, and the mismatches are counted. A comparison of the results is presented in Table 3.

The SPC features present an average F-measure of 74.05%, the APP features average 71.87%, the APP&SPC feature combination yields 70.16%, and the SELECTED features set averages 69.02%. The spectral-based feature set showed better performance than the appearance-based feature set because the spectral-based extraction technique can extract only the significant features. By contrast, the appearance-based feature extraction cannot exclude the unwanted noise component and causes an average performance decrease of 2.18%. The improvement effort by combining the appearance- and spectral-base features degraded the performance by 3.89%. Another improvement effort, selecting the top \( n \) features (where \( n = 50 \)), also yields inadequate results; it produces a performance decrease of by 5.04%.

Figure 11 shows the results of the F-measure, precision rate, and recall rate for the overall, demung-only, and saron-only note tracking. The spectral-based feature achieved the highest F-measure (83.79%), with 86.47% for the precision rate. It is interesting that the F-measure for demung-only note tracking is 83.79%, higher than the demung-only F-measure, which reaches only 63.89%. Demung and saron are the balungan instruments that are played by imitating the core melody, so using the saron-only note tracking is a promising technique for producing balungan gendhing notation. However, it cannot be concluded that saron-only note tracking is the best technique for producing balungan gendhing notation. Further investigation of both instruments is needed by increasing the amount of data testing that involves demung and saron.

Table 3 shows that, on average, the performance of the experiments with testing data containing the bonang instrument was substantially worse than the performance of other data testing. The performance worsened because of two or more events on the same (or adjacent) frequency and time, such as some of the bonang sound signals mixed with the saron or demung sound signals. Because of the overlapping of the first harmonic components, it is difficult to separate a mixture of sounds and extract the original features. However, there is opportunity to improve the performance, such as investigating the specific overtones corresponding to a gamelan instrument’s timbre.

4. CONCLUSION

In this paper, we presented methods for gamelan

| Data       | F-Measure (%) |
|------------|---------------|
|            | SPC | APP | APP&SPC | SELECTED |
|-------------|-----|-----|---------|----------|
| 1           | 71.43 | 64.97 | 58.46 | 66.67 |
| 2           | 85.33 | 82.91 | 95.33 | 90.65 |
| 3           | 92.11 | 98.06 | 81.63 | 82.71 |
| 4           | 69.73 | 55.22 | 51.72 | 47.46 |
| 5           | 96.97 | 95.54 | 94.70 | 96.97 |
| 6           | 55.21 | 71.83 | 57.83 | 54.67 |
| 7           | 84.62 | 86.61 | 86.82 | 84.62 |
| 8           | 89.80 | 100.00 | 94.74 | 92.11 |
| 9           | 84.24 | 82.55 | 83.57 | 80.70 |
| 10          | 46.76 | 51.80 | 59.65 | 49.64 |
| 11          | 89.58 | 63.91 | 77.81 | 69.42 |
| 12          | 77.09 | 44.44 | 38.36 | 52.92 |
| 13          | 56.25 | 56.95 | 54.92 | 47.24 |
| 14          | 45.24 | 51.84 | 53.04 | 50.62 |
| 15          | 66.38 | 71.39 | 63.81 | 68.87 |

Average: 74.05 | 71.87 | 70.16 | 69.02

Table 3 The F-measure result obtained with the test data set listed in Table 2. The SELECTED feature set is part of the appearance (APP) and spectral (SPC) based features set.
A novel fast feature extraction technique for face recognition,” *IEEE Trans. Syst. Man Cybern., Part B: Cybernetics*, **36**, 946–953 (2006).

[3] S.-K. Oh, S.-H. Yoo and W. Pedrycz, “Design of face recognition algorithm using PCA-LDA combined for hybrid data pre-processing and polynomial-based RBF neural networks: Design and its application,” *Expert Syst. Appl.*, **37**, 6404–6408 (2012).

[4] A. Abadpour and S. Kasaei, “Color PCA eigenimages and their application to compression and watermarking,” *Image Vision Comput.*, **26**, pp. 878–890 (2008).

[5] L. Malagón-Borja and O. Fuentes, “Object detection using image reconstruction with PCA,” *Image Vision Comput.*, **27**(12), pp. 2–9 (2009).

[6] H. de Paula, H. Yehia, J. Vasconcelos and M. Loureiro, “Representing timbre dynamics of a musical instrument: Comparison between GA and PCA,” *Proc. 14th IEEE Signal Process. Soc. Workshop Machine Learning for Signal Process.*, pp. 381–390 (2004).

[7] T. Kitahara, M. Goto and H. G. Okuno, “Pitch-dependent identification of musical instrument sounds,” *Appl. Intell.*, **23**, 267–275 (2005).

[8] J. C. Brown and B. Zhang, “Musical frequency tracking using the methods of conventional and “narrowed” autocorrelation,” *J. Acoust. Soc. Am.*, **89**, 2346–2354 (1991).

[9] Y. K. Suprapto, M. Hariadi and M. H. Purnomo, “Traditional music sound extraction based on spectral density model using adaptive cross-correlation for automatic transcription,” *IAENG Int. J. Comput. Sci.*, **38**(2), pp. 1–8 (2011).

[10] K. D. Martin and Y. E. Kim, “Musical instrument identification: A pattern-recognition approach,” *Proc. 136th Meet. Acoust. Soc. Am.*, ASA, pp. 1–11 (1998).

[11] S. Essid, G. Richard and B. David, “Instrument recognition in polyphonic music based on automatic taxonomies,” *IEEE Trans. Audio Speech Lang. Process.*, **14**, pp. 68–80 (2006).

[12] T. Taniguchi, M. Tohyama and K. Shirai, “Detection of speech and music based on spectral tracking,” *Speech Commun.*, **50**, 547–563 (2008).

[13] R. Chen, C.-F. Chan and H. C. So, “Model-based speech enhancement with improved spectral envelope estimation via dynamic tracking,” *IEEE Trans. Audio Speech Lang. Process.*, **20**, 1324–1336 (2012).

[14] E. Benetos and Y. Stylianou, “Auditory spectrum-based pitched instrument onset detection,” *IEEE Trans. Audio Speech Lang. Process.*, **18**, pp. 1968–1977 (2010).

[15] M. Ihara, S.-i. Maeda and S. Ishii, “Instrument identification in monophonic music using spectral information,” *Proc. IEEE Int. Symp. Sig. Process. Inf. Technol.*, pp. 595–599 (2007).

[16] V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[17] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” *Proc. 10th Eur. Conf. Machine Learning, ECML-98*, Chemnitz, C. Nédellec and C. Rouveiroi, Eds. (Springer Verlag, Heidelberg, 1998), pp. 137–142.

[18] I. Kotsia and I. Pitas, “Real time facial expression recognition from image sequences using support vector machines,” *Proc. IEEE Int. Conf. Image Process.*, ICIP2005, Vol. 2, pp. 966–9 (2005).

[19] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Mach. Learn.*, **46**, pp. 389–422 (2002).

[20] G. Costantini, M. Todisco and G. Saggio, “A new method for musical onset detection in polyphonic piano music,” *Proc. 14th WSEAS Int. Conf. Computers*, Stevens Point, Wisconsin.

**REFERENCES**

[1] L. E. Miller and F. Lieberman, “Lou Harrison and the American Gamelan,” *Am. Music*, **17**, 146–178 (1999).

[2] W. Zuo, D. Zhang, J. Yang and K. Wang, “BDPCA plus LDA: instrument sound recognition using the appearance and spectral-based features, as well as support vector machines as classifiers. The spectral-based features performed better than the appearance-based features and also better than the combination of appearance and spectral. However, the proposed method still cannot handle the condition where the “facial image” of two gamelan instrument sounds overlap each other, or in other words, if both signals have the same time and frequency.

Future work will include extending our method to deal with the segmentation stability, extracting the features from overtones, and applying several techniques to reveal hidden signals, such as cross-correlating and extracting useful features to overcome the overlap condition. It is also important to deal with a variety of gamelan instruments, including the kendang and the wooden xylophone.
USA, ICCOMP’10, Vol. 2, pp. 545–548 (2010).

[21] J. P. Bello, “Towards the automated analysis of simple polyphonic music: A knowledge-based approach,” Ph.D. Dissertation (Univ. of London, Queen Mary, 2003).

[22] A. Klapuri, “Signal processing methods for the automatic transcription of music,” Ph.D. Dissertation (Tampere University of Technology, 2004).

[23] J. R. B. Blázquez and J. P. de Leon Vázquez, “Estimation of the instantaneous amplitude and the instantaneous frequency of audio signals using complex wavelets,” Signal Process., 3093–3109 (2010).

[24] W. Surjodiningrat, Tone Measurements of Outstanding Javanese Gamelan in Jogjakarta and Surakarta, 2nd ed., (Gajah Mada University Press, Yogyakarta, 1993).

[25] G. Hu and D. Wang, “Auditory segmentation based on onset and offset analysis,” Trans. Audio Speech Lang., 15, 396–405 (2007).

[26] J. P. Bello, C. Duxbury, M. Davies and M. Sandler, “On the use of phase and energy for musical onset detection in the complex domain,” IEEE Signal Process. Lett., 11, 553–556 (2004).

[27] A. Tjahyanto, Y. K. Suprapto, M. H. Purnomo and D. P. Wulandari, “FFT-based features selection for Javanese music note and instrument identification using Support Vector Machines,” Proc. IEEE Int. Conf. Comput. Sci. Autom. Eng., CSAE., Zhangjiakou, China, pp. 439–443 (2012).

[28] C.-W. Hsu, C.-C. Chang and C.-J. Lin, A Practical Guide to Support Vector Classification (National Taiwan University, Taipei, 2003).

[29] I. Jolliffe, “Rotation of principal components: choice of normalization constraints,” J. Appl. Stat., 22, 29–35 (1995).

[30] G. Tzanetakis and P. R. Cook, “Musical genre classification of audio signals,” IEEE Trans. Speech Audio Process., 10, 293–302 (2002).

[31] L. T. DeCarlo, “On the meaning and use of kurtosis,” Psychol. Methods, 2, 292–307 (1997).

[32] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. (Morgan Kaufman, San Francisco, 2005).

[33] Sumarsam, Cultural Interaction and Musical Development in Central Java (The University of Chicago Press, Chicago, 1992, 1995).

[34] Y. K. Suprapto, I.-K. E. Purnama, M. Hariadi and M. H. Purnomo, “Sound Modeling of Javanese Traditional Music Instrument,” Proc. Int. Conf. Instrum. Commun. Inf. Technol. Biomed. Eng., ICICI-BME, Bandung: ITB, pp. 1–6 (2009).