Supervised Music Summarization for Human-Intuitive Highlight Identification

Jaesung Lee, Jihae Yoon, Seongwon Lee, and Dae-Won Kim

***Abstract*—With the great demand of automatic music analysis, Music Summarization that aims to determine the most representative segment in given music, paid much attention in music information retrieval field. In this paper, we propose a new approach to music summarization. Our goal is to identify the segment that listeners actually recognized as the most representative or memorable one. The strategy of proposed approach is to learn the relationship between acoustic features and information annotated by human instead of selecting a segment based on the self-structure of given music clip. Our prediction model can identify the most representative part considerably, and the experimental results also show that the proposed approach has significant potential to music summarization.**

***Index Terms*—Human-intuitive, music summarization, supervised learning, prediction system.**

I. Introduction

Recently, explosive number of music is published almost every day. To find music clips that users want to listen fast from massive candidates, the technique that provides an ability to browse music corpus quickly has paid much attention from music information retrieval field [1]. In this situation, Music Summarization (or Music thumbnailing) that aims to determine the most representative segment of a given music recording automatically, was considered one of the most suitable technique, because summarized music is able to deliver theme, and mood to the listeners shortly [2]-[4].

According to the notion of summary for music, several musical parts can be a good summarization candidates [4], [5]. In a past decade, most researches focused on identifying the repeated segment of music based on acoustic feature such as rhythm and tempo [3]. Although previous researches have shown that most listeners tend to perceive the repeated segment as the most representative one, it was not clearly concluded that the most representative segment is the repeated one [2], [4]. Thus, in this paper, we assume that listeners are able to identify the most representative segment even though it is not repeated in the music clip.

Our goal is to identify the segment that listeners actually recognized as the most representative or memorable one. In this approach, the most representative segment is determined by the listener directly instead of internal structure of music. Therefore, the most representative segment should be supervised by the human. Based on the annotated information, automatic human-intuitive music summarization can be achieved after an algorithm learns the relation between acoustic features and annotated information. To accomplish this task successfully, it should be investigated that the algorithm is able to identify the most representative segment accurately. In this paper, as a pilot study, we propose a new approach to find human-intuitively identified segment for music summarization.

Manuscript received October 30, 2015; revised January 17, 2016. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2014R1A1A2059032).

The authors are with the Chung-Ang University, Seoul, Korea (e-mail: jslee.cau@gmail.com, jihae.cau@gmail.com, seongwon.cau@gmail.com, dwkim@cau.ac.kr).

II. Related Works

Recent approaches for summarizing given music clips was tried to identify the most representative segment based on the internal structure of music itself, and the detection of repeated segment based on the rhythmic characteristics or chromagram based detection method was regarded as the most popular way to obtain most representative segment [1], [3], [4], [6]. According to this approach, the Self Similarity matrix (SSM) in terms of acoustic characteristics of given music clip firstly, and then repeated segment is chosen based on SSM.

M. Muller *et al*. proposed music summarization method based on repetitions covering large parts of given music clip [3]. In this research, they introduced a fitness function that assigns the description power for entire recording to each segment. The experimental results demonstrated that the advantage of this method is the stability to acoustic variations.

N. Jiang and M. Muller introduced a summarization technique considering multiple acoustic characteristics of music clip; verse and refrain section [4]. In this study, two approaches, independent and joint summarization based on two acoustic information were considered and compared. In previous work, they also considered the music summarization in the viewpoint of computational efficiency [1]. To accelerate the summarization process, they proposed multi-level sampling, segment resolution level control, and fitness value estimation strategy.

B. McFee and D. P. W. Ellis investigated a supervised learning method that optimizes acoustic features considering temporal characteristics of music clips [6]. To achieve the summarized segments, they proposed a latent structural repetition feature. The empirical results showed that the proposed method summarizes music clips efficiently.

T. Endrjukaite and Y. Kiyoki presented a music summarization method based on the tunes similarity because repetitions in tunes gave significant impression to the listeners [2]. To describe the tune information for calculating SSM, they developed a new descriptor based on the frequency.

M. Cooper and J. Foote presented several procedures for

music summarization [12]. To find the most representative part, they assume that the most representative part is mostly similar to entire part in an average sense. They built a SSM, then summed this matrix over the support value of each segment to the significance of being representative part, and then evaluate the similarity of that segment against the entire part. Finally, a part that gives the best similarity value is selected to be the most representative part.

A. Bartsch and H. Wakefield presented a system for producing representative samples [13]. This system searches for structural redundancy within a given music clip in terms of the chorus and refrain. For representation, they mapped the structure of given music clip into chroma-based representation, which is extracted from the information based on the cyclic attribute of pitch perception. After empirical experiments was conducted, they argued that their system effectively identify the repetitive part from given music clip.

III. Proposed Approach

A. Motivation

One limitation of previous approach is that there is no injection of the human perception, i.e. the most representative segment was determined by music itself with the absence of users or listeners. However, in most cases with consideration of music recommendation service or system, the success is determined by the listeners or users. Thus, the service should meet the user's satisfaction or expectation because users will not use this system if it does not deliver demanded output. Therefore, it is natural that the system incorporates the prior knowledge to its internal procedure.

To deliver qualified output, it is necessary that the system designers know about the characteristics or factors of leading good output, i.e. most representative segment of given music. Unfortunately, it is not clearly defined what is the *most representative* segment, or what characteristics of music signal leads to the perception of *most representative* even though users naturally determine the most representative segment of given music. In this case, it is possible to avoid the difficulty by using the user's annotation to the most representative segment. After the system learns the relation between input information and the most representative segment reliably, we may define our target based on the input information, e.g. acoustic feature values. In addition, the automation on the music summarization can also be achieved by applying the obtained model to unseen music clip.

Because our approach highly depends on the learnt model, the success of analysis affected by the quality of obtained model. If the model outputs accurate output, then it is able to assume that this model well-identified the relation between acoustic features and annotation to the most representative segment. Therefore, validating the quality of obtained model is necessary before conducting detailed analysis. To obtain a learnt model, we need three components; music corpus, users' annotation to determine the most representative segment, and learning algorithm or prediction system.

B. Data Gathering and Annotation

The procedure of constructing the data set is given below. First, each music clip is segmented with a fixed duration. Second, musical properties were extracted from segmented music clips. Feature extraction was performed by using short-term Fourier transform, heuristic musical property detectors, and so on. Those signal transformation methods are applied to each music clip, and then musical properties were converted from audio signals to numerical values. Second step of forming the data set is to collect supervised information that annotates the most representative segment.

We collected 55 music clips from popular three genres; Ballad, Electronic, and Hip-Hop. Two annotators have listened those music clips separately and mark the starting time of the most representative part of given music. Then each music clip was segmented by every 10 seconds with 5 seconds overlap. As a result, 2,597 candidate segments were obtained. We choose the most representative segment corresponding to marked starting time. In this procedure, multiple segments can be chosen as the most representative segment if it is found several times in entire music. Finally, 690 segments among 2,597 segments were identified as the most representative segment for each genre. Table I represents the detailed information about annotation and corresponding results.

TABLE I: Annotation Result of Music Clips

Genre	No. of Non-Representative Segments	No. of Representative Segments	Total
Ballad	530	176	706
Electronic	682	303	985
Hip-Hop	695	211	906
Total	1907	690	2,597

For each music segment, we used the MIR toolbox that offers integrated set of functions to extract musical audio features [7]. The extracted features fall into six types: dynamics, fluctuation, rhythm, spectral, timbre, and tonal features. At the ends, we obtained our data set which is composed of 2,597 patterns and 365 features.

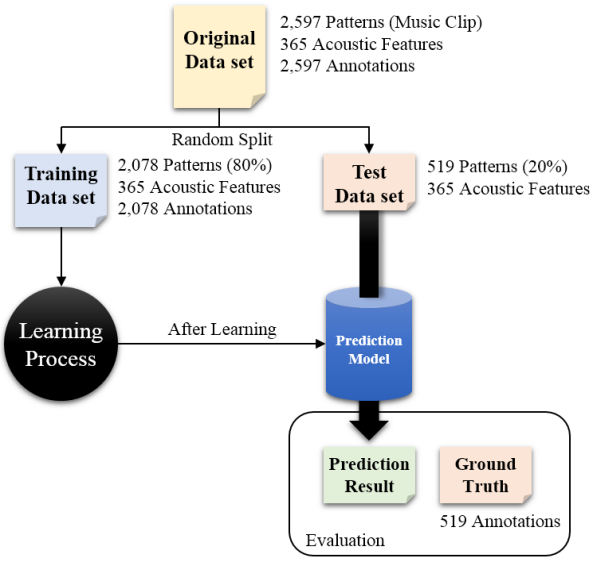

Fig. 1. Procedure of Blind Cross-validation.

C. Obtaining Reliable Prediction Performance

After the data set is composed, we need to imply a model

based on the obtained data set. Although manual analysis may be favorable with regard to the given problem, it becomes labor-intensive task if the data set is composed of many patterns and features. In this case, we can consider well-known expert systems or prediction systems whose prediction process is done by data-driven model and is easy-to-understandable to human expert. Last a few decades, many prediction systems were publicized according to the constraints they focused on [8]. Therefore, it might be necessary that choosing a prediction system well-suited to given data set. Typically, the choice is made by simulating and comparing the prediction results of candidate systems.

Another concern to obtain reliable model may rise by the over-fitting; the prediction system creates too specific model for given data set so that it loses the generality to unseen patterns. Therefore, too optimistic prediction performance may be obtained, and inaccurate prediction can be performed in the practical situation. To avoid this, many cross-validation schemes were devised. In this paper, we employed one of the most widely-used cross-validation scheme that is known as the blind test [9]; the data set was separated into two parts and then one part is only used for training (or learning) process, and the other part is used for testing the prediction performance. Fig. 1 shows the detailed simulation procedure based on the blind test. Because the test part never affects to the training process, it is able to avoid the problem caused by the over-fitting.

IV. Experimental Results

A. Experimental Settings

We conducted experiments related to the performance of the proposed approach for music summarization. To validate the suitability of our approach, we choose conventional Naïve Bayes classifier after testing several classifiers; Naïve Bayes classifier gave the best prediction performance for our data set. The Naïve Bayes classifier creates classification model for six classes in our data set. Then this prediction model is used to output the likelihood value of each class for given test pattern. Therefore, these likelihood value can be regarded as the confidence to each class. We discretized our data set using an equal-width interval scheme to apply the Naïve Bayes classifier [10]. To select number of bins, we observed the prediction accuracy by changing the number of bins to two, three, four, and five. Among them, the classifier gives better classification performance when the number of bins is set to three. As a result, we mapped each numerical feature into a categorical feature with three bins for our experiments.

To obtain more realistic prediction accuracy, we conducted a hold-out cross-validation for our experiments; 80% of the patterns in a given data set were randomly chosen as training set for the prediction process, and the remaining 20% of the patterns were used as test set to obtain the prediction accuracy value to be reported. The simulation was repeated 100 times, and the average value was taken to represent the prediction accuracy. Finally, we employed the Area Under the Receiver Operating Characteristics (AUROC) to consider the quality of the predicted classes; AUROC values can be drawn from zero to one with high value indicates the better prediction.

B. Preliminary Analysis

In many cases, it is quite preferable that the distribution of given data set is known to the researcher because it provides important insights and expectations for further analysis, especially when we are unaware or clear definition of target concept. However, when the data set is composed of too many features, the direct plotting is impossible. To circumvent this situation, we applied Principal Component Analysis (PCA) on our data set, and plotted the distribution of music clips on the reduced acoustic feature space projected by PCA [11].

Fig. 2 shows PCA projection results of 2,597 segments; horizontal axis represents the first principal component and vertical axis indicates the second principal component given by PCA. In the figure, cross mark, rectangle, and circle indicate Ballad, Hip-Hop, and Electronic music segments, respectively. Experimental result shows that Electronic music segments are well-separated from other music segments of two genre. Although Ballad music segments and Hip-Hop music segments are slightly overlapped in the acoustic feature space, most of them are separated from each other; Ballad music segments are distributed in the bottom region whereas Hip-Hop music segments are distributed in the top region.

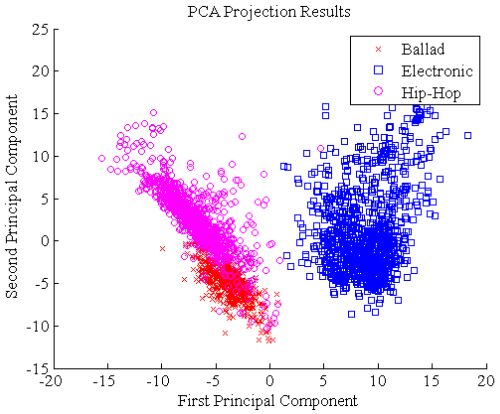

Fig. 2. PCA projection results of music clips.

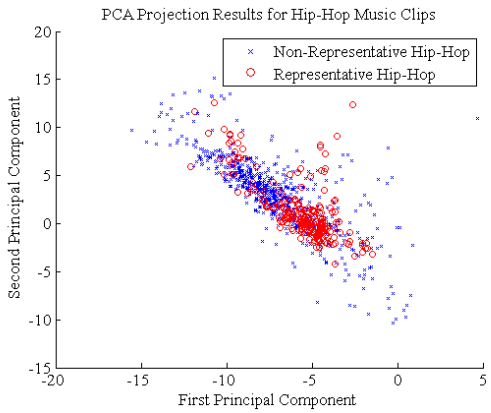

Fig. 3. PCA projection results of hip-hop music segments.

For detailed analysis, we plotted Hip-Hop music segments as shown in Fig. 3. In the figure, cross mark and circle indicate non-representative segments and representative segments from Hip-Hop music segments, respectively. The figure shows that most non-representative segments are distributed in the top-left region whereas most representative segments are distributed in the bottom-right region. Similar to the tendency observed from Fig. 2, the figure indicates that

representative segments and its counterparts take significantly different positions on the acoustic feature space. For the experiments of other segments from remaining two genre, we observed similar tendency of distribution.

C. Prediction Performance

To validate the accuracy of obtained model, we observed the prediction accuracy by using AUROC curve. Table II shows the experimental results of the representative segments prediction. In the table, the values represent AUROC curve value and corresponding standard derivation. Experimental results show the prediction was performed very accurately because AUROC of six types more than 0.9. Thus, Table II indicates that the prediction model successfully captured the relation between acoustic features and the annotation to the most representative segments.

TABLE II: EXPERIMENTAL RESULTS OF SEGMENT PREDICTION

AUROC	Ballad	Electronic	Hip-Hop
Non-Representative Segments	0.95 (0.008)	0.98 (0.005)	0.95 (0.008)
Representative Segments	0.94 (0.010)	0.96 (0.008)	0.92 (0.013)

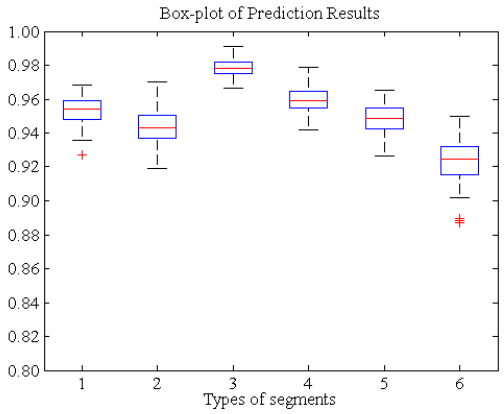

Fig. 4. Box-plot of prediction performance distribution.

TABLE III: MEANING OF EACH TYPES IN FIG. 4

Class	Meaning
1	Non-representative Ballad Segments
2	Representative Ballad Segments
3	Non-representative Electronic Segments
4	Representative Electronic Segments
5	Non-representative Hip-Hop Segments
6	Representative Hip-Hop Segments

In our simulation for evaluating the prediction performance, the prediction accuracy can be different according to each iteration because the training data set and test data set is separated randomly for each simulation. However, in most cases, it will be better that the prediction system outputs good prediction results stably. To validate this aspect, we represent the distribution of prediction performance. Fig. 4 shows the box-plot of the AUROC values and their variances for each type of segments. Because of the space limitation, we represent each type by using integer value in the figure, and the meaning of each class explain in Table III.

In Fig. 4, for example, the results of prediction on non-representative Electronic segments show a maximum prediction performance of 0.99 and a minimum prediction performance of 0.97, as well as median, upper quartile and lower quartile value of 0.98. Experimental results show that the minimum prediction performance more than 0.9 for all six types of segment, indicating the obtained model outputs good prediction results with significantly stable manner.

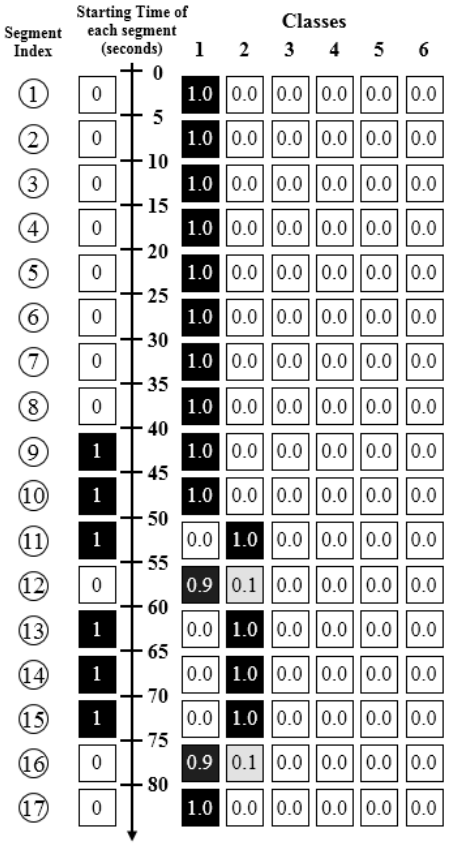

Fig. 5. Graphical illustration of highlight prediction.

To show how the proposed system identifies the most representative segment from given music clip, we illustrate the identification results of an example music clip included in our data set, by representing annotated highlight segment and corresponding predicted significance value according to each segment. Fig. 5 shows the identification results for a Ballad music clip entitled as "forget-me-not" by Korean artist Sae-Byul Park. Because of the space limitation, we represented the identification results of 0 to 90 seconds duration, resulting in 17 segments due to 5 seconds overlap; the index of each segment is represented in the left most side of the figure. In the figure, the most representative segment annotated by human expert, is placed in the next of segment index. Each square represents corresponding segment (10 seconds duration) and the value in the square indicates the ground truth; if the value is 1, then it indicates corresponding segment is annotated as the most representative segment in given music clip. In this example, the segments 9, 10, 11, 13, 14, and 15 are annotated as the most representative segments. Next, the vertical axis represents the starting time of each segment in seconds. In the right side of the vertical axis, each of six column corresponds to each class defined in Table III. The value in each square of the same row represents predicted significance value for each class. For example, the system

assigned the significance values of 0.9, 0.1, 0.0, 0.0, 0.0 and 0.0 to Class 1, 2, 3, 4, 5, and 6 for the segment 12, indicating this segment is identified as Non-representative segment of Ballad music clip. It should be noted that the sum of significance values for a segment is always 1.0, whereby the class of each segment can also be determined by choosing the class with the largest significance value among six values.

The experimental result shows that our proposed system assigned zero significance value to Class 3, 4, 5, and 6 along with most segments because of different genre; the genre of input music clips is Ballad whereas Class 3, 4, 5, and 6 are correspond to Non-representative / Representative segments of Electronic and Hip Hop genre. Therefore, the proposed system accurately distinguished the genre of given music clip. The figure also shows that the proposed system was identified Non-representative segments for given music clip accurately because the system assigns the maximum significance value of 1.0 to most Non-representative segments. In the example music clip, there are two representative parts; segments 9, 10, and 11, and segments 13, 14, and 15. Experimental result shows the proposed system appropriately assigned the significance values to the segments 13, 14, and 15, whereas the significance values for segments 9 and 10 are incorrectly assigned. A possible reason for this results is that the system assigns the significance value based on the information extracted from other segments in the data set; the segments 9 and 10 can be a representative segment within just this music clip, however, with the consideration of all the other reference segments in the data set, the segments 9 and 10 are unlikely to be a representative segment for Ballad.

V. Discussion

Because this is a pilot study for our new approach to music summarization, we have several issues that have to be complemented. One possible consideration is about the number of genre in our data set. In our study, the experimental results indicated that the most representative segment of Ballad, Electronic, and Hip-Hop can be predicted accurately by the obtained model. Specifically, the model tends to predict segments of Electronic genre more accurately than segments of other genre. A possible reason for this results is the distribution of Ballad and Hip-Hop; they are distributed closely in the acoustic feature space as shown in Fig. 2. Thus, the prediction model may be confused Ballad and Hip-Hop genre. Although three genre considered in our study are quite popular music genre, there remain many other music clips in the real-world that can be assigned to other genre. To validate the availability and practicality of our approach against diverse music style, more music clips should be gathered and included to our data set. After that, we can observe more reliable performance of our approach.

The second concern is about the number of annotators. In our study, two annotators participated to the annotation of the most representative segments. Although they are likely to agree to annotate the same segment as the most representative segment in our study, it should be validated whether the general consensus about the most representative segment can be obtained or not when massive annotators participated to the annotation. In this way, a possible misjudgment on the most representative segment can be circumvented, resulting in more accurate model for supervised highlight detection.

The third concern is related to the annotation of each music segment. In our study, each music clip was segmented by 10 seconds duration with 5 seconds overlap. As a result, the starting time of representative part may be mismatched to the segment, resulting in possible identification error. This issue may be achieved by introducing real-valued annotation, denoting the significance of representativeness based on the portion of actual highlight duration in the segment. Moreover, averaging the significance values given by participants is also considerable to obtain a more reliable annotation. In either ways, we expect that more accurate highlight detection model can be obtained because detailed annotation can prevent possible confusion due to 0/1 discrimination.

The last concern is related to the analysis on the prediction model. Although the proposed approach has shown its potential to predict human-intuitive music summarization, it was still not identified what acoustic features are correlated to the perception to the most representative segment. To achieve this, the prediction function in the obtained model should be more analyzed in detail. In order to discard irrelevant acoustic features and obtain an easy-to-interpret model, introducing an effective feature selection method for our data set is highly desirable. By doing this, we can identify the meaning of the most representative segment, bridging the human perception and computational model based on the acoustic features.

VI. Conclusion

In this paper, we proposed a new approach to summarize music clips under the concept of human-intuitively chosen most representative segment. As a pilot study, we examined the potential of proposed approach and the experimental results showed the intuitive music segment that is perceived by users can be automatically and accurately predicted.

References

[1] N. Jiang and M. Muller, "Towards efficient audio thumbnailing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 4-9, 2014, pp. 5192-5196.

[2] T. Endrjukaite and Y. Kiyoki, "Music similarity analysis through repetitions and instantaneous frequency spectrum," in *Proc. 5th International Conference of Signal Processing Systems*, Sydney, Australia, December 6-7, 2013, pp. 170-176.

[3] M. Muller, N. Jiang, and P. Grosche, "A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 531-543, 2013.

[4] N. Jiang and M. Muller, "Estimating double thumbnails for music recordings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Australia, April 19-24, 2015, pp. 146-150

[5] M.-B. Chung and I.-J. Ko, "Representative melodies retrieval using digital signal processing of audio," in *Proc. International Conference on Hybrid Information Technology*, Cheju Island, Korea, November 9-11, 2006, pp. 185-190.

[6] B. McFee and D. P. W. Ellis, "Learning to segment songs with ordinal linear discriminant analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 4-9, 2014, pp. 5192-5196.

[7] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in *Proc. International Conference on Music Information Retrieval*, Vienna, Austria, September 23-27, 2007, pp. 237-244.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newletter*, vol. 11, no. 1, pp. 10-18, 2009.
[9] J. Lee and D.-W. Kim, "Mutual information-based multi-label feature selection using interaction information," *Expert Systems with Applications*, vol. 42, no. 4, pp. 2013-2025, 2015.
[10] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Proc. 12th International Conference on Machine Learning*, Tahoe City, USA 1995, pp. 194-202,.
[11] M. Zhang, J. Pena, and V. Robles, "Feature selection for multi-label Naïve Bayes classification," *Information Sciences*, vol. 179, no. 19, pp. 3218-3229, 2009.
[12] M. Cooper and J. Foote, "Automatic music summarization via similarity aanlysis," in *Proc. 3rd International Conference on Music Information Retrieval*, Paris, France, October 13-17, 2002.
[13] M. Bartsch and G. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96-104, 2005.

Jaesung Lee received M.S. and Ph.D. in computer science from Chung-Ang University, Korea in 2009 and 2013, respectively. He currently participates in Post-doctoral course at Chung-Ang University. His research interest includes biomedical informatics and affective computing. In theoretical domain, he also studies classification, feature selection, and multi-label learning with information theory.

Jihae Yoon is currently a master student of the School of Computer Science and Engineering in Chung-Ang University, Seoul, Korea. She started her position from 2015.

Seongwon Lee is currently a master student of the School of Computer Science and Engineering in Chung-Ang University, Seoul, Korea. He started his position from 2015.

Dae-Won Kim is currently a professor in the School of Computer Science and Engineering, Chung-Ang Univ. in Seoul, Korea. Prior to coming to CAU, he did his postdoc, Ph.D., M.S. at KAIST, and the B.S. at Kyungpook Nat'l. Univ., Korea. His research interest includes advanced data mining algorithms with innovative applications to bioinformatics, music emotion recognition, educational data mining, affective computing, and robot interaction.