

MODELING DYNAMIC PATTERNS FOR EMOTIONAL CONTENT IN MUSIC

Yonatan Vaizman

Edmond & Lily Safra Center
for Brain Sciences, ICNC.
Hebrew University
Jerusalem Israel

yonatan.vaizman@mail.huji.ac.il

Roni Y. Granot

Musicology Dept.
Hebrew University
Jerusalem Israel

rgranot@huji.013.net.il

Gert Lanckriet

Electrical & Computer
Engineering Dept.
University of California
San Diego

gert@ece.ucsd.edu

ABSTRACT

Emotional content is a major component in music. It has long been a research topic of interest to discover the acoustic patterns in the music that carry that emotional information, and enable performers to communicate emotional messages to listeners. Previous works looked in the audio signal for local cues, most of which assume monophonic music, and their statistics over time. Here, we used generic audio features, that can be calculated for any audio signal, and focused on the progression of these features through time, investigating how informative the dynamics of the audio is for emotional content. Our data is comprised of piano and vocal improvisations of musically trained performers, instructed to convey 4 categorical emotions. We applied Dynamic Texture Mixture (DTM), that models both the instantaneous sound qualities and their dynamics, and demonstrated the strength of the model. We further showed that once taking the dynamics into account even highly reduced versions of the generic audio features carry a substantial amount of information about the emotional content. Finally, we demonstrate how interpreting the parameters of the trained models can yield interesting cognitive suggestions.

1. INTRODUCTION

There is a general agreement that music (especially instrumental music) lacks clear semantic information but conveys rich emotional content. As a form of non semantic communication, musical performers are able to convey emotional messages through the sound and listeners are able to interpret the sound and figure out the emotional intention of the performer. What are the patterns in the musical signal itself that enable this communication? The properties

of the musical content that are responsible for carrying this emotional information have long been the subject of interest and research. In previous computational research that analyzed emotions expressed in music performance, some works looked for local acoustic cues, such as *notes per second*, *articulation degree*, etc., that are present in the sound and may play a significant role in conveying the emotional message [1,2]. Statistics of these cues over time were calculated and were usually used to train a discriminative model. Calculations of these local cues from raw audio data usually rely on intermediate signal processing algorithms to detect note onsets and other events, and these intermediate calculations may introduce assumptions, errors and bias. In addition, such cues are often defined for monophonic music, and are sometimes even designed for specific instruments. While such analysis methods may be very useful for musical training and acquiring performance skills of conveying emotions, they tend to be very specific. Other works avoid this problem by using generic audio features, such as MFCCs or other spectral features. Such generic audio features are defined in a more straight forward way than sophisticated local cues, and don't require intermediate signal processing calculations. Although these features may not describe certain perceptual properties that the local cues try to capture, presumably they will be more robust. In addition, generic audio features don't assume anything on the signal, and can be applied to any audio signal, even if it contains polyphonic music and even multiple instruments. Such audio content will be a serious obstacle for the local cues approach. Several systems that participated in the MIREX evaluation apply the same audio features for different MIR tasks [3,4]. In those systems running average and standard deviations of time varying audio features were taken, but same as in the local cues approach, the complete dynamics of the audio wasn't used. Such methods disregard the order of time points and assumes they're independent.

In the presented work, we suggest an approach that addresses both the issues of specificity and dynamics. We apply generic audio features (Mel frequency spectrum) to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

© 2011 International Society for Music Information Retrieval.

	<i>happy</i>	<i>sad</i>	<i>angry</i>	<i>fearful</i>	total
piano	8	7	7	6	28
vocal	12	12	12	12	48
total	20	19	19	18	76

Table 1. Distribution of the recordings over emotions and instrument.

overcome the specificity problem. The dynamics issue is resolved by using Dynamic Texture Mixture (DTM) [6]. DTM was designed to model both the instantaneous properties of a time series signal, and their dynamics. This model enables us to capture important information that resides in the course of change of the audio through time, which is missed when assuming independence among time points.

Similar dynamic systems were used by Schmidt and Kim [5] to model the time-varying distribution of emotional state (in 1 sec intervals). Here we regard each musical instance (improvisation of about 30 seconds) as conveying a single emotional message (described simply by an emotional adjective), and we apply the dynamic system on the lower level of the time-varying audio features themselves.

DTMs and Gaussian Mixture Models (GMMs) have been applied to music information retrieval systems, including semantic tags of emotional categories annotated by listeners as being relevant for popular songs [7, 8], but not yet applied to audio recordings specifically created to convey emotional content. The data in the presented work has recordings of improvisations by musically trained performers instructed to convey specific emotions.

2. METHODS

2.1 Data

Our data set is comprised of 76 audio recordings of musically trained performers (2 pianists and 2 vocalists, 1 female and 1 male in each category). For each recording the performer was instructed to improvise a short musical segment that will *convey to listeners in a clear manner a single emotion*, one from the set of $\{happy, sad, angry, fearful\}$. These emotional instructions were used as the ground truth labels for the recordings (3 judges verified that the appropriate emotions are expressed. Future analyses will also regard ratings from a larger group of listeners as labels). These improvisations clearly rely, in part, on well entrenched cultural musical norms and even clichés. Thus we obtained a relatively wide variety of acoustic manifestations for each emotional category, which presumably capture the various strategies and aspects of how these specific emotions can be conveyed in Western music. The distribution of recordings over emotions and instrument is detailed in Table 1. The median duration for a recording was 24 seconds.

2.2 Audio features

Mel spectrum features were collected: for each time frame Discrete Fourier Transform was calculated and the energy of the frequency components was integrated in overlapping frequency bins, in a Mel scale, and the $10\log_{10}$ of the bins' energies were saved. Similarly to [7] we used 34 Mel frequency bins (partitioning the band from 20Hz to the Nyquist frequency, 11kHz, to 34 Mel-scaled bins), and used half overlapping time frames of 2048 samples (after re-sampling the audio data to 22,050Hz, this results in a feature vector every 46msec).

2.3 Modeling the dynamics

In order to model the dynamics of acoustic properties of the music, we applied the Dynamic Texture Mixture (DTM) model. DTM was previously used to model dynamic textures of video [6] and of audio [7]. A *Dynamic Texture (DT)* is a generative model for a time sequence of observed features (e.g. the acoustic features collected for each short time frame), that assumes that the observed feature vector $y(t)$ was generated as a linear transformation (plus additive Gaussian noise) over an internal state - a hidden vector variable $x(t)$ (possibly in a much smaller dimension than the observed feature vector). It also assumes the dynamics of the hidden variable is a Linear Dynamic System, driven by additive Gaussian zero-mean noise: the state of the hidden variable at any time point $x(t)$ depends only on its state in the previous time point $x(t-1)$, and the dependency is linear.

$$\begin{cases} x_t &= Ax_{t-1} + v_t \\ y_t &= Cx_t + w_t \end{cases} \quad (1)$$

Where v_t and w_t are both random normal variables (drawn independently for each t). A *DTM* is a mixture of DTs, each having a different relative weight. The DTM models the generation of an audio instance (a song) as follows: for each segment of the song first select a DT out of the mixture (according to the weights of the DTs), and then generate the observed acoustic features of the segment from the selected DT.

Since this is a generative model, we can calculate the likelihood of a song (or of a collection of songs) given a DTM. This facilitates the ranking of songs according to their likelihood of being generated by a given DTM or the ranking of different DTMs according to the likelihood of a song of being generated by them. The parameters of a DTM can be learned from training data, using an iterative Expectation Maximization algorithm tailored for learning DTMs (EM-DTM) [6].

For each of the 4 emotions (*happy, sad, angry* and *fearful*), sequences of 125 consecutive feature vectors were collected (in order to get many feature sequences to train on,

we used overlapping sequences, with hop of 15 feature vectors from sequence to sequence) from all the recordings in the training set that were associated with the emotion, and a DTM to represent that emotion was trained over these sequences. Since each feature vector represented a time frame of about 46msec, the resulting sequences represented segments of about 5.7 seconds. The median number of sequences collected for a recording was 26. We used DTMs with 4 components (4 DTs), and with dimension of 7 for the hidden variable x (unless the observed features were in a lower dimension).

2.4 Performance evaluation

In order to evaluate the success of the acoustic features to represent the required information regarding the emotional content, and the success of the model to capture the relevant acoustic patterns for the emotional content, we used information retrieval framework and performance measures: After training 4 emotion DTMs on the training set, a test set with unseen recordings was analyzed. For each recording the 4 emotions were ranked according to the likelihood of that recording given the 4 emotion DTMs, and annotation of 1 emotion (the one with highest likelihood) was given to the recording. For each emotion, the test recordings were ranked according to their likelihood given the DTM of the emotion, as a retrieval task. Comparing the machine’s annotation and retrieval to the ground truth emotion labels of the test recording, 3 annotation measures and 2 retrieval measures were calculated, in a similar manner to [7]: *precision* (portion of the ground truth labeled instances out of the machine-annotated instances), *recall* (portion of the machine-annotated instances out of the ground truth labeled instances), *f-measure* (balance measure between precision and recall), mean average precision -*MAP* (average precision over different thresholds of “how many of the top-ranked instances to retrieve”) and area under ROC curve -*AROC* (area under the tradeoff curve of true-positive rate vs. false-positive rate for the retrieval task, area of 0.5 being chance and area of 1 being maximum possible). Each of the 5 measures was calculated for each emotion, and then averaged over emotions.

To estimate these measures over general unseen data, 10-fold cross validation scheme was used. For each partition, 4 emotion-DTMs were trained over 9/10 of the recordings, and the 5 measures were calculated over the remaining 1/10 of the recordings. In each partition control performance measures (chance level) were approximated by repeatedly (400 times) generating random uniform values (instead of the likelihood values actually calculated with the trained models) and feeding them to the annotation-retrieval system, for the test set. Mean and standard deviation over repetitions were collected as reference for assessment of quality of the actual performance scores. Approximated p-values were then calculated to each of the 5 measures, as the prob-

	precision	recall	F	MAP	AROC
score	0.6446	0.6500	0.6000	0.8099	0.8692
chance	0.25	0.25	0.22	0.44	0.50
p-val	0.09	0.04	0.06	0.02	0.02

Table 2. Annotation and retrieval results for basic features.

ability of getting a higher score under the null hypothesis, meaning with random values (assuming a normal distribution with the mean and standard deviation that we collected for the random values). Finally we averaged over the 10 folds the 5 performance measures, as well as 5 chance level scores and 5 p-values for our scores. The partition to 10 folds was semi random, making sure each fold contained instances from all 4 emotional categories, and all experiments were done using the same partitioning to 10 folds, in order for the comparison to be consistent.

3. EXPERIMENTS AND RESULTS

3.1 Experiment 1 - basic

The system was applied to the basic features as described above. The results of the cross validation are presented in Table 2. In the basic experiment, the results demonstrate that the DTM model manages to capture the important acoustic patterns for the communication of emotion.

3.2 Experiment 2 - power dynamics

In order to investigate the role of the power dynamics, two complementary manipulations over the features were performed:

Ex2.1: flattening the power. For each recording, all the Mel spectra vectors were normalized to have the same constant total power, but within each vector, the original ratios among the frequency bins were preserved. This manipulation filters out the power dynamics (in time scales larger than 46msec), and keeps all the rest of the information stored in the original features (melody, timbre, etc.).

Ex2.2: keeping only the power dynamics. For each recording and for each time point, instead of keeping 34 coefficients, only 1 coefficient is kept - the total power of the time frame (in log scale). This manipulation preserves only the power dynamics, and filters out the rest of the sound properties. Since the observed features in every time frame were then only 1 dimensional, the dimension of the hidden variable x was also reduced to 1, resulting in a linear dynamic system that is almost degenerate (since the transition matrix A is simply a scalar), and relies more on the driving noise.

Ex2.3: not modeling dynamics. As control, using the same features as in Ex2.2 we applied a GMM model that assumes independent time frames, to see if we can still capture the remained relevant information about the emotions,

		precision	recall	F	MAP	AROC
Ex2.1	score	0.6134	0.6375	0.5775	0.7627	0.8429
	p-val	0.07	0.04	0.05	0.04	0.02
Ex2.2	score	0.4287	0.4625	0.3801	0.5935	0.6879
	p-val	0.19	0.14	0.18	0.16	0.14
Ex2.3	score	0.2931	0.3125	0.2638	0.5536	0.6454
	p-val	0.44	0.4	0.42	0.29	0.25

Table 3. Results for power manipulations.

while disregarding the dynamics. The only dependency left among time frames was the 1st and 2nd time derivatives (delta and acceleration) of the feature vector (of the power scalar, in this experiment) that were augmented, so the feature vector here was 3 dimensional (for time t : $power(t)$, $delta(t) = power(t+1) - power(t)$ and $acceleration(t) = delta(t+1) - delta(t)$). For training we used the hierarchical EM algorithm for GMM (HEM-GMM), as described in [8]. We used 4 components (4 Gaussians) for each model (each GMM), and restricted to diagonal covariance matrices.

Results are presented in Table 3 (the reference chance levels, which appear in Table 2, are the same in all experiments). Ex2.1 demonstrates that most of the information about the conveyed emotion is retained even without the gross dynamics of the power (keeping in mind that some finer power dynamics can be expressed inside each time frame, in the lower frequency bins). Although this may suggest that the gross power dynamics doesn't carry much information about the emotions, Ex2.2 shows the contrary: after reducing the features to only the power dynamics, the scores remain fairly high (although, as expected for a 1 dimensional time function, some decrease in performance is evident). The results show that the power dynamics does carry useful information about the emotional content. The control done in Ex2.3 shows that GMM got very poor scores for the 3 annotation performance measures, and relatively poorer results than DTM (Ex2.2) for all measures. It is quite expected that when reducing the features to only the power, treating the time frames as independent will yield insufficient information about the emotions. The gap between the results of Ex2.2 and Ex2.3 shows the added value of taking into account the dynamics of the acoustical properties (when even 1st and 2nd time derivatives are not enough).

3.3 Experiment 3 - avoiding frequency correlations

When acoustical instruments (or human voice) are playing, the harmonic structure has correlations between the fundamental frequencies and their higher harmonics, resulting in correlation between the dynamics of different frequency bins, and suggesting redundancy when all these frequency bins are specified. The DTM model deals with this redundancy by trying to find a lower representation in the hidden state x that generates the observed vectors (by linear transforma-

		precision	recall	F	MAP	AROC
Ex3.1	score	0.5288	0.5667	0.4847	0.7331	0.7969
	p-val	0.23	0.13	0.19	0.13	0.1
Ex3.2	score	0.4701	0.4375	0.3648	0.6213	0.6546
	p-val	0.14	0.16	0.21	0.16	0.21
Ex3.3	score	0.1718	0.175	0.1339	0.4354	0.5425
	p-val	0.67	0.65	0.69	0.51	0.4

Table 4. Results for keeping part of the spectrum.

tion with the observation matrix C - the principal components of the observed features) [7]. We wanted to reduce the observed features prior to summarizing the whole spectrum and, in a way, to overlook the correlations among frequencies. For this purpose we examined limiting our view to only part of the spectrum. We focused on two opposite extremes of the spectrum captured by the original features:

Ex3.1: 6400Hz-11025Hz (Nyquist frequency). Keeping only the last 6 frequency bins of each time frame. Such frequency band is likely to contain resonating frequencies to the fundamental frequencies of the melody being played (or voiced). When calculating mean over all time frames in all instances in the data set, these 6 bins carry only 0.036 of the power (not log power) of the spectrum.

Ex3.2: 0Hz-275Hz. Keeping only the first 3 frequency bins of each time frame. For part of the time frames this frequency band may be below the present fundamental frequency of the tones being played. These 3 bins carry (in average) 0.25 of the power of the spectrum. For both Ex3.1 and Ex3.2 we used dimension of 3 for the hidden variable x . These extreme bands probably behave differently for piano and for vocal and interesting insights can later be raised by performing similar experiments separately for instruments.

Ex3.3: not modeling dynamics. Similar to the control done in Ex2.3, we applied the GMM model to the features used in Ex3.2, plus 1st and 2nd time derivatives.

Results are presented in Table 4. Ex3 demonstrates that in both extremes of the spectrum, there are small frequency bands that still carry a fair amount of information about the conveyed emotions (performance is still relatively far from chance level). The control in Ex3.3 that, again, shows poor results with the GMM (performance being about chance level or worse), affirms that the remained relevant information lies mostly in the dynamics.

3.4 Experiment 4 - melodic structure

Next we aimed to examine the affect of the melodic dynamics on the conveyed emotions. Since it is neither simple nor accurate to determine the notes that were played, especially for polyphonic music (such as our piano recordings), we chose to define a more accurate property that hopefully will be more robust: the dynamics of the strongest frequency bin. We cannot claim to describe the perceived melody (or the played notes) with this property (since pitch percep-

	precision	recall	F	MAP	AROC
score	0.4322	0.4375	0.3852	0.58	0.6863
p-val	0.23	0.2	0.21	0.2	0.16

Table 5. Results for keeping only strongest frequency bin.

tion or production is more complex than just the strongest frequency present, and since the piano music has multiple tones played simultaneously). However this property is easily computed and can be used as a surrogate marker for the melodic progression. For this experiment, in each time frame only the strongest bin remained active and the power of all the other frequency bins was nullified. Furthermore, to get rid of the power dynamics, the power of all time frames was set to be constant, so the only remaining information was the identity of the activated bin in each time frame. The dimension of the hidden variable x was set to 1. Results are shown in Table 5. Although the features were reduced to a large extent, a predictive ability is still present.

3.5 Interpreting the trained models

After validating that DTMs can capture important acoustic patterns for emotional content, we wanted to understand the differences between different trained emotion models that enabled the system to discriminate. Using a generative model is suitable to describe the process of production: the performers that want to convey some emotion and apply an appropriate generative strategy to create their resulting sound. In order to get insight about the different generative strategies, one needs to compare the learned parameters of the trained models. For this purpose we retrained 4 emotion DTMs over the entire data set, for our different experiments.

The main component that describes the dynamics of the system in a DT is the transition matrix A . If the system were a deterministic linear dynamic system, without the additive noise, this transition matrix would tell both the destination of the state of the system x and the way it will take to get there. The eigenvectors of A describe the different modes of the system - different patterns of activating the observed features. The eigenvalues of A (complex numbers) indicate the course of progress of the different modes (patterns) of the system: while having an eigenvalue with magnitude larger than 1 results in the state of the system diverging, having an eigenvalue with magnitude of 1 results in the state converging to either a stable state or stable limit cycle determined by the eigenvector of that value, and having all eigenvalues with magnitudes smaller than 1 results in a system that strives to converge to the zero vector state (if there is no additive noise to reactivate the modes). The magnitude of an eigenvalue indicates the intensity or stability of this mode (how slowly this mode will decay or how much anti-mode noise needs to be added to this mode in order to silence it). The angle of the eigenvalue indicates the

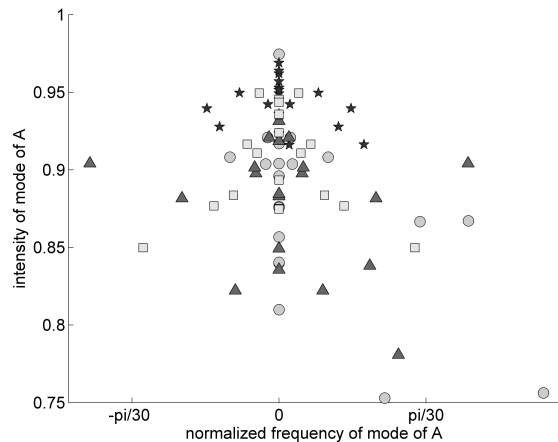


Figure 1. Eigenvalues for using the basic features (Ex1). Each different shape represents 20 eigenvalues of transition matrices from DTs of a different emotion DTM (5 largest eigenvalues from 4 DTs per emotion-DTM). *happy* - circle, *sad* - star, *angry* - triangle, *fearful* - square.

normalized frequency of the mode - if an eigenvalue has a large angle its mode will oscillate and modulate its pattern in a fast period, returning to the original modulation pattern (only with smaller magnitude) after only few time frames. The maximal normalized frequency will be π , making the mode change to its exact negative in each consecutive time frame. We examined the eigenvalues of the different DTs of the different emotion DTMs, and presented their magnitudes (intensity) and angles (frequency).

In both conditions presented in Figure 1 and Figure 2 there is a clear concentration of the eigenvalues of the *sad* model (marked with star) with relatively high intensities and low frequencies (in absolute value). This can be interpreted as a general strategy (either conscious or subliminal) of the performers to convey sadness using stable and slowly modulated acoustic patterns. On the opposite, the *happy* and *angry* models (marked by circle and triangle, respectively) include many modes with smaller intensities and higher frequencies, suggesting strategies that include fast repetitions of acoustic patterns (high frequencies) and easy switching from one dominating pattern to another (the low magnitudes mean that little noise is sufficient to shake off these modes and activate different modes).

Such conclusions should be taken with a grain of salt. We should remember the system also has additive noise. In addition, in order to adequately generalize these results, much larger data sets, with many performers, should be used. However, such analyses may help to focus future research on certain aspects of production of music for emotional communication.

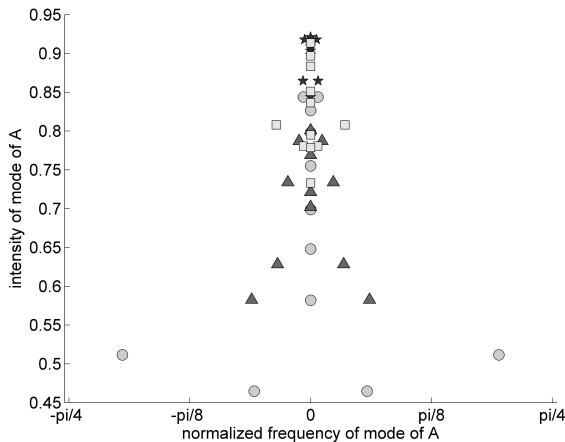


Figure 2. Eigenvalues for keeping only higher frequency band (6 kHz-11 kHz. Ex3.2).

4. DISCUSSION

Investigating the dynamics of generic acoustic features in musical audio can reveal important components of the music, and specifically for emotional content. Generic acoustic features can be informative for various melodic, harmonic, rhythmic and instrumental content of music, and here we demonstrated their successful usage for both monophonic and polyphonic music. We have shown that even highly reduced audio features, such as the power, can still retain much of the emotional message, when taking into account the time progression of the property. Interestingly, complementary manipulations to reduce the audio features (“flattening the power” vs. “keeping only the power dynamics”) both kept a discriminative ability, suggesting that the information about the emotional intention carried by separate components of the sound is not simply additive, but rather having redundancy. One should remember, though, that it might require few dimensions of features to discriminate 4 emotions, but possibly require more detailed features, when discriminating more emotions and emotional subtleties.

Future research using similar methods should be applied over more general musical data, with multiple instruments, to find general dynamic patterns that convey different emotions. It may be interesting to investigate the critical time resolutions that show dynamics that is relevant for emotional content (perhaps taking sequences of more than 125 time frames will reveal slower informative patterns). Experiments with larger data will enable investigating differences in strategies, in informative frequency bands, redundancy patterns and other aspects, among different emotions. Another interesting direction is to use trained generative models to synthesize new audio instances. This is not a simple

challenge, but even if the resulting sounds will not be intelligible or natural sounding, they may still have an effect of conveying emotions, and concordance between the emotion of the generated audio and that of the generating model will be another convincing argument that the model captures important acoustic patterns for emotional communication.

5. ACKNOWLEDGEMENTS

Dr. Avi Gilboa, Dr. Ehud Bodner, Liza Bekker and Nori Jacobi took part in the collection of the data. Special thanks to Emanuele Coviello for guidance with DTM. Gert Lanckriet acknowledges support from Yahoo! Inc. and NSF Grant IIS-1054960

6. REFERENCES

- [1] L. Mion, and G. De Poli: “Score-Independent Audio Features for Description of Music Expression,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol.16, No. 2, pp. 458–466, 2008.
- [2] A. Friberg, E. Schoonderwaldt, P. Juslin, and R. Bresin: “Automatic Real-Time Extraction of Musical Expression,” *Proceedings of the International Symposium Computer Music Conference*, pp. 365–367, 2000.
- [3] G. Tzanetakis: “Marsyas Submission to MIREX 2009,” *MIREX 2009*.
- [4] G. Peeters: “a Generic Training and Classification System for MIREX08 Classification Tasks: Audio Music Mood, Audio Genre, Audio Artist and Audio Tag,” *MIREX 2008*.
- [5] E. M. Schmidt and Y. E. Kim: “Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering,” in *Proceedings of the 2010 IEEE International Conference on Machine Learning and Applications*, 2010.
- [6] A. B. Chan, and N. Vasconcelos: “Modeling, Clustering and Segmenting Video with Mixtures of Dynamic Textures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.30, No. 5, pp. 909–926, 2008.
- [7] E. Coviello, A. B. Chan, and G. Lanckriet: “Time Series Models for Semantic Music Annotation,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol.19, No. 5, pp. 1343–1359, 2011.
- [8] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet: “Semantic Annotation and Retrieval of Music and Sound Effects,” *IEEE Transactions on Audio, Speech and Language Processing*, Vol.16, No. B2, pp. 467–476, 2008.