Automatic Music Video Generation Based on Emotion-Oriented Pseudo Song Prediction and Matching

Jen-Chun Lin
Academia Sinica
Taipei, Taiwan
jenchunlin@gmail.com

Wen-Li Wei
Academia Sinica
Taipei, Taiwan
lilijinjin@gmail.com

Hsin-Min Wang
Academia Sinica
Taipei, Taiwan
whm@iis.sinica.edu.tw

ABSTRACT
The main difficulty in automatic music video (MV) generation lies in how to match two different media (i.e., video and music). This paper proposes a novel content-based MV generation system based on emotion-oriented pseudo song prediction and matching. We use a multi-task deep neural network (MDNN) to jointly learn the relationship among music, video, and emotion from an emotion-annotated MV corpus. Given a queried video, the MDNN is applied to predict the acoustic (music) features from the visual (video) features, i.e., the pseudo song corresponding to the video. Then, the pseudo acoustic (music) features are matched with the acoustic (music) features of each music track in the music collection according to a pseudo-song-based deep similarity matching (PDSM) metric given by another deep neural network (DNN) trained on the acoustic and pseudo acoustic features of the positive (official), less-positive (artificial), and negative (artificial) MV examples. The results of objective and subjective experiments demonstrate that the proposed pseudo-song-based framework performs well and can generate appealing MVs with better viewing and listening experiences.

Keywords
Automatic music video generation, cross-modal media retrieval, deep neural networks.

1. INTRODUCTION
With the prevalence of mobile devices, video is widely used to record memorable moments of daily events such as wedding, graduation, and birthday parties. Websites such as YouTube or Vimeo have furthered the phenomenon as sharing becomes ever easy. In addition, people enjoy listening to music to release their emotions. In psychology, it is argued that a musical experience may evoke emotions when a listener conjures up images of things and events that have never occurred, in the absence of any episodic memory from a previous event in time [1]. Thus, music and video are often accompanied to complement each other to enhance emotional resonance in music videos (MVs), movies, and television programs. Accompanying a user-generated video (UGV) with music can enhance the entertaining quality and emotional resonance, and thus is desirable. For example, a wedding video accompanying with romantic music can enhance a sweet atmosphere. Nevertheless, to select good music for a video, music professionals are required. With the rapid growth of music collections, matching a video with suitable music becomes ever difficult. The advent of automatic MV generation systems is foreseeable.

In response to this trend, machine-aided MV composition has been studied in the past decade [2–8]. However, the performance of previous systems is usually limited, because most of them only consider the relationship between the low-level acoustic and visual features without considering any semantic constraints [2–4]. Since different semantics, e.g., emotion such as happy and sad, usually contain distinct MV properties, there is difficulty in establishing a direct relationship between the music and video modalities from the low-level features without considering such semantics. Moreover, there is a so-called semantic gap between the low-level acoustic (or visual) features and the high-level human perception. To narrow the gap, motivated by the recent development in affective computing of multimedia signals, some research has begun to map the low-level acoustic and visual features into an emotional space [5–8]. A music-accompanied video composed in this way is attractive, as the perception of emotion naturally occurs in video watching and music listening. However, most of the existing studies for automatic MV generation [5–8] model the relationship between the low-level acoustic (or visual features) and the emotion labels separately, whereas ignoring the correlation between music and video. Since the music and video contents in a professionally edited official music video (OMV) are always highly synchronized and carefully composed to match each other in terms of emotional storytelling, without considering the relationship between the music and video modalities in automatic MV generation may still result in bad viewing experiences.

Our idea to jointly handle the aforementioned problems is inspired by the recent computational models of the brain [9,10], in particular the memory-prediction framework [10], which emphasizes the notion of multisensory spatiotemporal predictions. For example, based on the input from one sense, e.g., vision, the brain can predict the current and future events in other senses, e.g., hearing. Similar findings have also been reported in psychology and cognitive science. For example, it has been suggested in [11] that visual information has a predictive role in processing audio information. Driven by these findings, we propose a novel automatic MV generation framework based on emotion-oriented pseudo song prediction and matching, as shown in Figure 1. Given a queried video, a shot change detection method is first used to segment the queried video into several video shots. For each video shot, a multi-task deep neural network (MDNN) [12,13], which is trained by jointly learning the relationship among acoustic (music) features, visual (video) features, and

MM'16, October 15–19, 2016, Amsterdam, Netherlands
 ISBN 978-1-4503-3603-1/16/10 …$15.00
DOI: http://dx.doi.org/10.1145/2964284.2967245

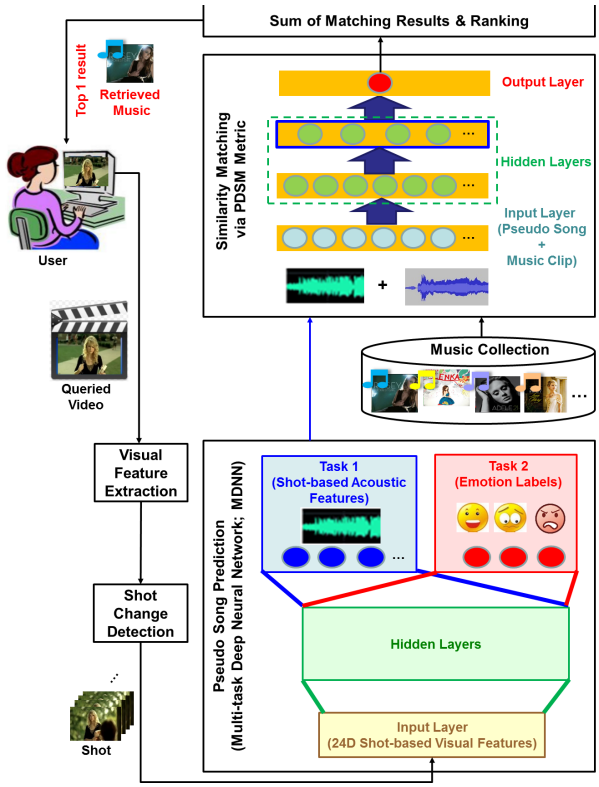

Figure 1. The MV generation framework based on emotion-oriented pseudo song prediction and matching.

emotion labels from an emotion-annotated OMV corpus, is adopted to predict the pseudo acoustic (music) features from the visual (video) features. For MV generation, a pseudo-song-based deep similarity matching (PDSM) metric is then applied to evaluate the similarity between the acoustic features of a music clip and the pseudo acoustic features of a shot of the queried video. The PDSM metric is realized by another deep neural network (DNN) trained on the positive (official), less-positive (artificial), and negative (artificial) MV examples. To the best of our knowledge, this is the first attempt to apply the multi-task deep neural network and pseudo song prediction and matching in automatic MV generation. The experimental results demonstrate that the proposed framework outperforms an existing system in both subjective and objective evaluations.

The remainder of this paper is organized as follows. Previous research on video soundtrack recommendation and MV generation is reviewed in Section 2. The methodology, including video shot change detection, pseudo song prediction, and pseudo-song-based deep similarity learning and matching, is described in Section 3. Finally, the experimental results are presented in Section 4, and conclusions are made in Section 5.

2. RELATED WORK

In this section, we briefly review the progress on video soundtrack recommendation and MV generation in the recent five years. Kuo et al. [4] employed multi-modal latent semantic analysis to learn the co-occurrence relationship between the low-level acoustic and visual features, such as Mel-frequency cepstral coefficients, loudness, spectral centroid, color, and motion, from the music and video contents for video soundtrack recommendation. To narrow the semantic gap between the low-level features and the high-level human perception, Wang et al. [5] proposed an acoustic-visual emotion Gaussians (AVEG) model to respectively map the acoustic features and the visual features into the same valence-arousal (VA) emotional space to measure the distance between a music clip and a video clip for MV generation. Shah et al. [6] employed a support vector machine (SVM) to model the categorical emotion, such as sweet, funny, and sad, from the acoustic, visual, and geographic features for video soundtrack recommendation. Lin et al. [7] adopted an emotional temporal course model (ETCM) to respectively model the temporal structure of emotional expression of music and video and a stream matching method to measure the similarity between the recognized emotional temporal phase sequences of music and video for MV generation. They further proposed an emotion-oriented deep similarity matching (EDSM) metric to measure the similarity between the recognized emotional temporal phase sequences [8].

3. METHODOLOGY

In the proposed MV generation system, as shown in Figure 1, shot change detection is first applied to segment a queried video into several video shots. For each video shot, a MDNN is used to predict the acoustic (music) features from the visual (video) features, called pseudo song prediction. Finally, a PDSM metric is used to match music and video based on the acoustic features.

3.1 Video Shot Change Detection

A queried video usually contains thousands of image frames. Consider that a video shot is usually captured by a single camera action, and that there are no significant content changes between successive frames in a shot [14], it would be more efficient to conduct pseudo song prediction and matching at the shot-level instead of the frame-level. In this study, we use 5-color themes [15] (a kind of dominant color representation) as features to segment the video into shots. The shot boundaries are determined according to the color theme difference CTD_i between two adjacent image frames I_{i-1} and I_i calculated as

$$CTD_i = \prod_{k=1}^{5} \prod_{j=1}^{5} CT_i(k) - CT_{i-1}(j), \quad (1)$$

where $CT_i(k)$ denotes the k-th color theme in the i-th frame and $CT_{i-1}(j)$ denotes the j-th color theme in the i-1-th frame. All the frames with $CTD_i \neq 0$ are considered shot boundaries. Figure 2 shows an example of the shot change detection results and the corresponding color themes. After shot change detection, the shot-based visual (or acoustic) features can be constructed by statistics of the component frame-based visual (or acoustic) features. If shot change detection fails to detect any boundaries, the whole video is used to predict a pseudo song.

3.2 Pseudo Song Prediction via Multi-Task Deep Neural Network

Multi-task learning [16] aims at improving the generalization performance of a learning task by jointly learning multiple related tasks together. It has been found that if the tasks are related and share some internal representation, then through joint learning, they can transfer knowledge to one another. The common internal representation learned in this way helps the models generalize better for the future unseen data. Consequently, for pseudo song prediction, we adopt the MDNN [12,13] to predict the acoustic (music) features from the visual (video) features by jointly

Figure 2. Shot change detection results of the OMV "love story" by Taylor Swift.

learning the relationship among the acoustic (music) features, visual (video) features, and emotion labels from an emotion-annotated OMV corpus.

Assume that there are K tasks $T \equiv \{T_1, T_2 ..., T_K\}$ to learn under the MDNN framework. The MDNN model parameters are represented by $\Lambda \equiv \{\lambda_0\} \cup \{\lambda_1, \lambda_2, ..., \lambda_K\}$, where λ_0 consists of model parameters shared by all tasks and λ_k consists of model parameters specific to task T_k. In this study, λ_0 represents the shared weights from all hidden layers, whereas λ_k represents the weights associated with the task-specific output layer of T_k. Without loss of generality, T_1 will always be taken as the primary task, and the rest are the secondary (or extra) tasks. The objective function ε for training is formulated as the weighted sum of the error functions of all the tasks as follows,

$$\varepsilon(D, \Lambda) = \sum_{x \in D} \left(\sum_{k=1}^{K} \beta_k \varepsilon_k \left(x; \lambda_0, \lambda_k \right) \right), \quad (2)$$

where ε_k and β_k are the error function and weight of task T_k subject to $\sum_{k=1}^{K} \beta_k = 1$, x is an input vector, and D is the whole set of training vectors for all tasks. After training, only the model parameters associated with the primary task T_1 (i.e., λ_0 and λ_1) are needed, and those of the secondary task(s) can be discarded.

In this study, we use emotion as the secondary task T_2 to learn the MDNN for predicting the acoustic features (the primary task T_1) from the input visual features x. The emotion can be regarded as the semantic constraint for MDNN learning, and is expected to be able to improve the prediction accuracy. β_k (k=1,2) is set to 0.5, and the error function ε_k of task T_k (k=1,2) is the sum of squared error as follows,

$$\varepsilon_k \left(x; \lambda_0, \lambda_k \right) = \sum_{i=1}^{N_k} \frac{1}{2} \times \left(d_i^{(k)} - s_i^{(k)} \right)^2, \quad (3)$$

where $d_i^{(k)}$ is the target value of the i-th output neuron for T_k, $s_i^{(k)}$ is the predicted value of the i-th output neuron for T_k, and N_k is the total number of output neurons for T_k. Specifically, x is the shot-based visual feature vector constructed from the component frame-based visual feature vectors, $d_i^{(1)}$ is the i-th element of the corresponding shot-based acoustic feature vector, $d_i^{(2)}$ is the emotion label, $s_i^{(1)}$ is the i-th element of the acoustic feature vector of the predicted pseudo song, while $s_i^{(2)}$ is the predicted emotion.

3.3 Pseudo-song-based Deep Similarity Learning and Matching for MV Generation

Recently, a similarity metric learning technique has been applied to the MV generation task [8]. The goal is to learn a flexible nonlinear similarity matching metric to alleviate the effect of emotion recognition errors in MV generation. Since inaccurate prediction of pseudo songs may degrade the performance of MV generation as well, a similarity metric capable of accommodating such pseudo song prediction errors is also desirable.

Again, we regard similarity learning as a regression learning problem. The goal is to learn a regression model (i.e., the PDSM metric) that can judge whether the acoustic features of a pseudo song and an arbitrary music clip of same length are similar. In PDSM metric learning, a DNN is adopted to learn the regression model based on a set of positive training examples v^{++}=(pseudo song, official music^{++}), less-positive training examples v^{+}=(pseudo song, official music^{+}), and negative training examples v^{-}=(pseudo song, official music^{-}) with labels y^{++}=3, y^{+}=2, and y^{-}=1, respectively. A positive training example is formed by the pseudo song and the music clip associated with a video shot of an OMV. A less-positive training example is constructed from the pseudo song of a video shot of an OMV and the music clip (same length as the pseudo song) of another OMV in the same VA emotional quadrant. A negative training example is constructed from the pseudo song of a video shot of an OMV and the music clip (same length as the pseudo song) of another OMV in a different VA emotional quadrant.

Denoting a training example v^{++}, v^{+} or v^{-} as v, we forward v layer-by-layer through a DNN to generate the representation of each layer, i.e., $v^{(1)}$,…, $v^{(L)}$. The l-th layer takes as input $v^{(l)}$ and uses a projection function to transform $v^{(l)}$ to $v^{(l+1)}$ as follows,

$$v^{(l+1)} = f^{(l)}(W^{(l)} v^{(l)} + b^{(l)}), \quad (4)$$

where $v^{(l)}$ and $v^{(l+1)}$ are the feature representation in the l-th and $l+1$-th layer, respectively; $W^{(l)}$ is a weight projection matrix; $b^{(l)}$ is a bias vector; and $f^{(l)}(.)$ is an activation function, which is a sigmoid function for l=1 to L-2, and a linear function for l=L-1. Given the label y, we use the sum of squared error as the loss function in the output layer:

$$\ell(v, y) = SSE(v^{(L)}, y). \quad (5)$$

The loss of the output layer will be back propagated to fine-tune the parameters W and b through the classical back-propagation method. Since the side-information (i.e., positive, less-positive or negative label) is considered for DNN to learn a nonlinear similarity matching metric, the resulting DNN regression model (i.e., the PDSM metric) is expected to alleviate the effect of errors in pseudo song prediction. The difference between the EDSM metric [8] and the proposed PDSM metric is that we regard similarity learning as a regression learning problem rather than a classification problem. By considering additionally the less-positive training examples, the learned PDSM metric should have a better generalization ability.

In the MV generation phase, given a queried video, the goal is to find a ranked list of music candidates for the query. Specifically, the queried video is paired with each music track from the target music database to form a testing pair. After video shot change detection, MDNN is applied to obtain the pseudo song for each video shot. Each paired music track is also divided into a sequence of music clips according to the time codes of the video shots. The PDSM metric is then applied to measure the similarity between the acoustic features of the pseudo song and the music clip, for each video shot. For each testing pair, the overall similarity score is summed from the similarity scores of all video shots. Finally, all the music tracks are ranked in

Table 1. Average ranking accuracy of the DEMV-matchmaker$_{com}$ and pseudo-song-based frameworks.

The Video Retrieving Music (V2M) Task	
DEMV-matchmaker$_{com}$ [8]	Pseudo-song-based MV Generation Framework
0.6519	0.7627

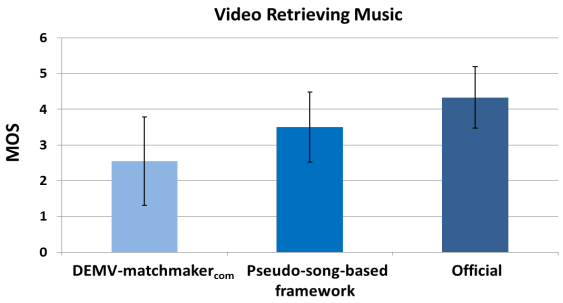

Figure 3. Results of the subjective MOS test.

descending order of scores, and the top one is regarded as the best recommendation for the queried video to generate the MV.

4. EXPERIMENTS

To evaluate the effectiveness of the proposed pseudo-song-based MV generation framework, we performed experiments on a set of OMVs downloaded from YouTube. 265 complete OMVs were collected, among which 65 OMVs downloaded according to the links provided in the DEAP database [17] were used to train the MDNN and the PDSM metric. Each OMV was assigned one (out of three) emotional quadrant based on the valence-arousal annotations provided in the DEAP database. The two emotional quadrants in the low arousal space were merged into one [7,8], since emotions mapped into the lower arousal space are difficult to differentiate [18]. The remaining 200 OMVs were used for testing.

For music, we used MIRToolbox to extract four types of frame-based acoustic features, namely dynamic, spectral, timbre, and tonal features [19,20]. In total, 46-dimensional acoustic features were extracted for each audio frame. Given a queried video, we extract the mean from the audio frames corresponding to a video shot as the 46-dimensional shot-based acoustic features. For video, the frame-based color themes and motion intensities were extracted as the 8-dimensional visual features [15,21]. The minimum, mean, and maximum values from the frame-based visual features in each video shot were extracted to generate 24-dimensional shot-based visual features. For the MDNN, there were 3 hidden layers, each with 230, 120, and 30 neurons, respectively. The size of mini-batch for the stochastic gradient descent algorithm was set to 20. For the PDSM metric, we used a DNN with 3 hidden layers, each with 230, 120, and 130 neurons, respectively. The size of mini-batch for the stochastic gradient descent algorithm was set to 1. For both DNNs, we applied random initialization for the weights, a constant learning rate of 0.05, and the L2 weight decay regularization to avoid over-fitting.

We compared the proposed pseudo-song-based MV generation framework with the state-of-the-art DEMV-matchmaker$_{com}$ framework [8]. In the experiments, the video of each testing OMV was used in turn to search for the best matched music from the music tracks of the 200 testing OMVs, and the one corresponding to the test video was regarded as the ground truth. The ranking accuracy [4] defined as

$$Ranking\ Accuracy = 1 - \frac{rank(g) - 1}{|C| + 1}, \quad (6)$$

was adopted as the objective performance measure, where *rank(g)* is the rank of the ground truth *g*, and $|C|$ is the total number of candidates in the music set ($|C|$=200 in this study). We reported the average ranking accuracy over the testing set.

The results in Table 1 demonstrate that the proposed pseudo-song-based MV generation framework outperforms DEMV-matchmaker$_{com}$. We believe that it is because DEMV-matchmaker$_{com}$ did not consider the relationship between music and video modalities in the respective emotion recognition model construction. It may lose information useful for music (or video) emotion recognition, since the music and video contents in an OMV are always highly synchronized and carefully composed to match each other in terms of emotional storytelling. Even DEMV-matchmaker$_{com}$ has integrated a similarity learning metric to alleviate the effect of emotion recognition errors, the performance is still limited. The multi-task deep neural network (i.e., MDNN) used in the pseudo-song-based MV generation framework can indeed model the relationship among the acoustic (music) features, visual (video) features, and emotion labels. Overall, the pseudo-song-based MV generation framework pushed ahead the rank of ground truth music by approximately 22 (i.e., the average ranking accuracy was improved from 0.6519 to 0.7627), compared to DEMV-matchmaker$_{com}$.

Subjective evaluation[1] in terms of 5-point mean opinion score (MOS) was conducted on 6 MV sets. Each MV set contains the original official MV (ground truth) and the MVs generated by DEMV-matchmaker$_{com}$ and the proposed pseudo-song-based MV generation frameworks. Each MV was evaluated by thirteen subjects. The average MOS over all MVs and subjects is shown in Figure 3. It is clear that the pseudo-song-based MV generation framework outperforms DEMV-matchmaker$_{com}$. The results reveal that modeling the relationship among music, video, and emotion can indeed generate more attractive MVs to enhance subjects' viewing and listening experiences. The results also show that the MOS of the MVs generated by the pseudo-song-based MV generation framework is quite close to that of the ground truth MVs. We feel very excited for this result.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a novel content-based MV generation system is proposed based on emotion-oriented pseudo song prediction and matching. The results of both subjective and objective evaluations have demonstrated that the proposed pseudo-song-based framework outperforms the state-of-the-art DEMV-matchmaker$_{com}$ framework, and can offer a satisfactory generated music video to enhance human viewing and listening experiences. Benefit from rich music video resources on the websites such as YouTube or Vimeo, developing an end-to-end deep neural network learning technique for automatic MV generation is desirable and will be studied in our future work.

6. ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant: NSC 102-2221-E-001-008-MY3.

[1] MOS results for individual MVs are available at https://sites.google.com/site/pseudosongpredictionmatching/

7. REFERENCES

[1] Juslin, P. N. and Västfjäll, D. Emotional responses to music: the need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31(5): 559–621, 2008.

[2] Hua, X.-S., Lu, L., and Zhang, H.-J. Automatic music video generation based on temporal pattern analysis. In *ACM MM*, 2004.

[3] Yoon, J.-C., Lee, I.-K., and Byun, S. Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools and Application*, 41(2): 197–214, 2009.

[4] Kuo, F.-F., Shan, M.-K., and Lee, S.-Y. Background music recommendation for video based on multimodal latent semantic analysis. In *ICME*, 2013.

[5] Wang, J.-C., Yang, Y.-H., Jhuo, I.-H., Lin, Y.-Y., and Wang, H.-M. The acousticvisual emotion Gaussians model for automatic generation of music video. In *ACM MM*, 2012.

[6] Shah, R. R., Yu, Y., and Zimmermann, R. ADVISOR–personalized video soundtrack recommendation by late fusion with heuristic rankings. In *ACM MM*, 2014.

[7] Lin, J.-C., Wei, W.-L., and Wang, H.-M. EMV-matchmaker: emotional temporal course modeling and matching for automatic music video generation. In *ACM MM*, 2015.

[8] Lin, J.-C., Wei, W.-L., and Wang, H.-M. DEMV-matchmaker: emotional temporal course representation and deep similarity matching for automatic music video generation. In *ICASSP*, 2016.

[9] Friston, K. The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*, 11(2): 127–138, 2010.

[10] Hawkins, J. and Blakeslee, S. *On intelligence*. Owl Books, 2005.

[11] Summerfield, Q. Some preliminaries to a comprehensive account of audio-visual speech perception. *Hearing by Eye: The Psychology of Lip Reading*, D. B. and C. R., Eds., 1987, 3–51.

[12] Chen, D. and Mak, Brian K.-W. Multitask learning of deep neural network for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7): 1172–1183, 2015.

[13] Xia, R. and Liu, Y. A multi-task learning framework for emotion recognition using 2D continuous space. *IEEE Transactions on Affective Computing*, 2015.

[14] Li, L., Zeng, X., Li, X., Hu, W., and Zhu, P. Video shot segmentation using graph-based dominant-set clustering. *ACM ICIMCS*, 2009.

[15] Wang, X., Jia, J., and Cai, L. Affective image adjustment with a single word. *The Visual Computer*, 29(11): 1121–1133, 2013.

[16] Caruana, R. Multitask learning. Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, USA, 1997.

[17] Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. DEAP: a database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1): 18–31, 2012.

[18] Soleymani, M., Kierkels, Joep J. M., Chanel, G., and Pun, T. A Bayesian framework for video affective representation. In *ACII*, 2009.

[19] Wang, J.-C., Yang, Y.-H., Wang, H.-M., and Jeng, S.-K. The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In *ACM MM*, 2012.

[20] Lartillot, O. and Toiviainen, P. A Matlab toolbox for musical feature extraction from audio. In *DAFx*, 2007.

[21] Chen, H.-W., Kuo, J.-H., Chu, W.-T., and Wu, J.-L. Action movies segmentation and summarization based on tempo analysis. In *MIR*, 2004.