EMV-matchmaker: Emotional Temporal Course Modeling and Matching for Automatic Music Video Generation

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

Wen-Li Wei
National Cheng Kung University
Tainan, Taiwan
lilijinjin@gmail.com

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

ABSTRACT

This paper presents a novel content-based emotion-oriented music video (MV) generation system, called EMV-matchmaker, which utilizes the emotional temporal phase sequence of the multimedia content as a bridge to connect music and video. Specifically, we adopt an emotional temporal course model (ETCM) to respectively learn the relationship between music and its emotional temporal phase sequence and the relationship between video and its emotional temporal phase sequence from an emotion-annotated MV corpus. Then, given a video clip (or a music clip), the visual (or acoustic) ETCM is applied to predict its emotional temporal phase sequence in a valence-arousal (VA) emotional space from the corresponding low-level visual (or acoustic) features. For MV generation, string matching is applied to measure the similarity between the emotional temporal phase sequences of video and music. The results of objective and subjective experiments demonstrate that EMV-matchmaker performs well and can generate appealing music videos that can enhance the viewing and listening experience.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.1 [**Information Interfaces and presentation**]: Multimedia Information Systems

Keywords

Automatic music video generation, computational emotion model, cross-modal media retrieval.

1. INTRODUCTION

With the prevalence of mobile devices, video is widely used to record the memorable moments of our daily events such as wedding, graduation, and birthday parties. Websites such as YouTube or Vimeo have furthered the phenomenon as sharing becomes 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 movie and television program. To enhance the entertaining quality and emotional resonance of user-generated videos (UGVs), accompanying a UGV with music is thus desirable. For example, a wedding video can accompany with romantic music to 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 an automatic music video (MV) generation system is foreseeable.

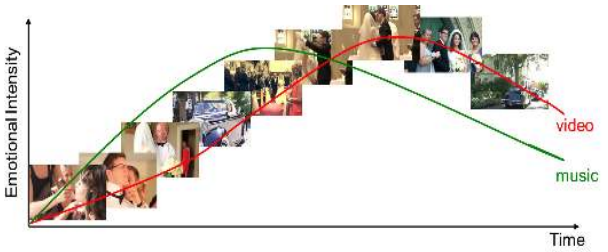

Figure 1. The mismatched temporal courses of emotional expression between music (green line) and video (red line); music reaches the emotional hot point (i.e., the vertex of the emotional curve) earlier than video.

In response to this trend, machine-aided automatic MV composition has been studied in the past decade [2–7]. However, the performance of existing systems is usually limited, because most of them only consider the relationship between the low-level acoustic features and visual features [3–5]. It is difficult to establish a direct relationship between the music and video modalities from low-level features. Moreover, there is a so-called semantic gap between the low-level acoustic (or visual) features and the high-level human perception. To narrow down such gap, motivated by the recent development in affective computing of multimedia signals, research has begun to map the low-level acoustic and visual features into the emotional space [6,7]. A music-accompanied video composed in this way is attractive, as the perception of emotion naturally occurs in video watching. However, most of the existing studies for MV generation [6,7] only model the relationship between the low-level features (i.e., acoustic or visual features) and the emotion labels, without considering the temporal course of emotional expression of music and video. Even the music and video are with the same emotional category, the nonsynchronous temporal courses of emotional expression may still result in bad viewing experience. Figure 1 shows an example of nonsynchronous temporal courses of emotional expression of video and music.

To handle the aforementioned problem, as shown in Figure 2, a novel emotion-oriented MV generation system called EMV-matchmaker is proposed in this paper. For a music (or video) clip,

MM '15, October 26-30, 2015, Brisbane, Australia
 ISBN 978-1-4503-3459-4/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2733373.2806359

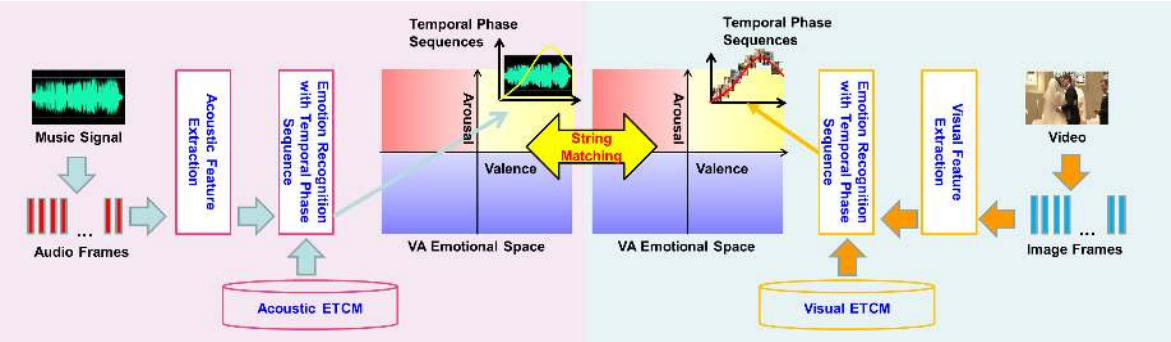

Figure 2. Illustration of the EMV-matchmaker framework.

an acoustic (or visual) emotional temporal course model (ETCM) is applied to predict its emotional temporal phase sequence in an emotional quadrant in the valence-arousal (VA) emotional space [8] from its low-level acoustic (or visual) features. For MV generation, the matching between music and video is based on whether they are located in the same emotional quadrant and the distance between their predicted emotional temporal phase sequences. The acoustic and visual ETCMs can be learned from an emotion-annotated MV corpus. To the best of our knowledge, this is the first attempt to consider the temporal structure of emotional expression for emotion-oriented MV generation.

2. RELATED WORK

Recent progress on video soundtrack recommendation and MV generation is overviewed here. Gillet et al. [3] employed the Pearson's correlation and mutual information to learn the correlation between the low-level acoustic and visual features, such as Mel-frequency cepstral coefficients, zero crossing rates, pitch, motion intensity, and color, from the music and video contents for MV generation; Similarly, Kuo et al. [5] employed multi-modal latent semantic analysis to learn the co-occurrence relationship between the low-level acoustic and visual features for video soundtrack recommendation, and Wu et al. [4] used multiple ranking canonical correlation analysis to learn such relationship for MV generation. To narrow down the semantic gap between the low-level features and the high-level human perception, Wang et al. [6] proposed an acoustic-visual emotion Gaussians (AVEG) model to respectively map the acoustic features and the visual features into the same VA emotional space to measure the distance between a music clip and a video clip for MV generation; Shah et al. [7] 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.

3. METHODOLOGY

In the proposed EMV-matchmaker system, as shown in Figure 2, an acoustic (or visual) ETCM is used to predict the emotional temporal phase sequence of a music (or video) clip in the VA emotional space. A string matching method is then used to match music and video based on the predicted emotional temporal phase sequences for MV generation.

3.1 Emotional Temporal Phase Sequence Prediction

The psychologist Ekman's research [9] demonstrated that the complete temporal course of an emotional expression can be divided into three sequential temporal phases, namely onset (application), apex (release), and offset (relaxation), considering the manner and intensity of the expression. To precisely model and predict the temporal course of an emotional expression, we adopt the ETCM, which was previously developed to model the temporal course of emotion from facial and speech expressions in [10,11]. We modified the ETCM to model the temporal course of emotional expression from MVs (including music and video contents). As shown in Figure 2, the EMV-matchmaker framework contains one acoustic ETCM and one visual ETCM for modeling music and video contents, respectively.

3.1.1 ETCM Derivation

In an ETCM, three emotional sub-states are defined to represent the temporal phases, namely onset, apex, and offset, of the emotional expression of a music clip (or a video clip), and a hidden Markov model (HMM) is used to model the temporal characteristics in an emotional sub-state.

Given an observation (i.e., acoustic or visual feature) sequence $O = o_1^T = o_1, o_2, ..., o_T$, the emotion recognition task is defined as selecting one among the three emotional quadrants $EQ \in \{EQ_1, EQ_2, EQ_3\}$ in the VA emotional space shown in Figure 2, i.e.,

$$EQ^* = \arg\max_{EQ} P(EQ \mid O). \quad (1)$$

For each emotional quadrant, $P(EQ \mid O)$ can be approximated as *a posteriori* probability of the best emotional sub-state (i.e., temporal phase) sequence $ES_{EQ} = es_{EQ}^1, es_{EQ}^2, ..., es_{EQ}^M$ as follows,

$$P(EQ \mid O) \approx \max_{ES_{EQ}} P(ES_{EQ} \mid O). \quad (2)$$

Therefore, the recognition problem is translated to finding out the emotional sub-state sequence that has the largest a *posteriori* probability over three emotional quadrants.

By using the Bayes' rule, the *a posteriori* probability $P(ES_{EQ} \mid O)$ can be decomposed as

$$P(ES_{EQ} \mid O) = P(O \mid ES_{EQ}) P(ES_{EQ}) / P(O), \quad (3)$$

where $P(O \mid ES_{EQ})$ is calculated by the corresponding emotional sub-state HMM sequence for the emotional quadrant EQ; $P(ES_{EQ}) = P(es_{EQ}^1, es_{EQ}^2, ..., es_{EQ}^M)$ is the *a priori* probability of the corresponding emotional sub-state sequence for the emotional quadrant EQ, which can be calculated according to a pre-defined grammar, as shown in Figure 3; $P(O)$ is identical for all possible emotional sub-state sequences, and thus can be omitted when (3) is applied in (1). Therefore, the task of emotion recognition with temporal phase sequence using ETCM can be expressed as

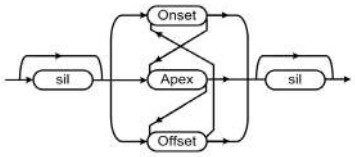

Figure 3. Recognition network based on the predefined grammar for characterizing an emotional quadrant expressed in a music or video clip.

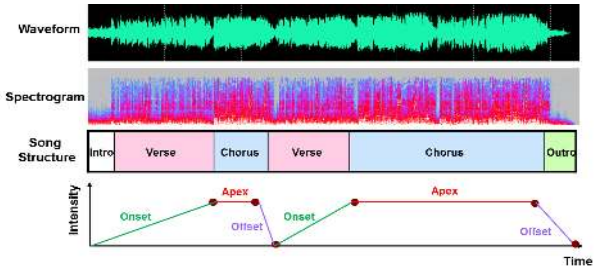

Figure 4. An example of annotated temporal phases of emotional expression of the song "Someone Like You" by Adele.

$$EQ^{*} = \arg\max_{EQ}\left[\max_{ES_{EQ}} P(O \mid ES_{EQ})P(ES_{EQ})\right]. \quad (4)$$

3.1.2 Emotional Temporal Phase Sequence Annotation and ETCM Training

We have collected a set of official music videos (OMVs) for training and testing. Since annotating the temporal course of emotional expression in an OMV is time-consuming, to accelerate the annotation process, two assumptions have been made. First, we assume that the temporal courses of emotional expressions for the music and video contents are the same in an OMV, since we believe that music and video are always highly synchronized and carefully composed to match each other in terms of emotion. Second, we assume that the temporal course of emotional expression in an OMV can be represented by a series of temporal phase (i.e., onset, apex and offset) combinations, since music is constructed based on a temporal structure consisting of intro, verse, chorus, bridge, and outro sections as well as optional repeats in order to over and over pave and express emotion. In our preliminary investigation, we observed that a complete emotional temporal phase sequence (i.e., onset-apex-offset) usually maps to a verse-chorus section in a song, as shown in Figure 4. We also observed that an apex phase is usually accompanied with more high frequency energy in the spectrogram. Based on these observations, to accelerate the annotation process, the temporal phases of each OMV in the training set were annotated according to the number of repetitions of verse-chorus sections by referring to the lyrics.

In acoustic ETCM training, for each emotional quadrant, we trained a set of HMMs (i.e., the acoustic emotional sub-state HMMs, including the onset HMM, the apex HMM, and the offset HMM) from the corresponding training OMVs together with the emotional temporal phase sequence annotation using the expectation-maximization (EM) algorithm. For each emotional sub-state HMM, a left-to-right HMM with three hidden states was used to model the emotional temporal characteristics. Similarly, we trained the visual emotional sub-state HMMs for constructing the visual ETCM. In addition, we also trained the sil HMMs (cf. Figure 3) to respectively absorb the black screen (for video) and the silence portion (for music) in the beginning and ending sections of an OMV, because these sections do not contain information of emotional expression. To permit the repetitions of emotional temporal phases in an OMV, an emotional temporal course grammar shown in Figure 3 was used to guide the recognition process by referring to the emotional temporal phases defined in [9]. All the temporal phase transition probabilities in the grammar were assumed uniformly distributed in this study.

3.2 MV Generation via String Matching

Given a queried video (or music) clip, the goal is to find a ranked list of music (or video) clips for the query. Specifically, the queried video (or music) clip is paired with each music (or video) clip from the target dataset to form a testing pair. The visual and acoustic ETCMs are then applied to predict the emotional temporal phase sequences $VES_{EQ} = (ves_{EQ,1}, ves_{EQ,2}, ..., ves_{EQ,T})$ and $MES_{EQ} = (mes_{EQ,1}, mes_{EQ,2}, ..., mes_{EQ,T})$ for the video and music clips by using (4), respectively. T is the length of the query. The similarity between the two emotional temporal phase sequences is evaluated by string matching as

$$S(VES_{EQ},\ MES_{EQ}) = \frac{1}{T}\sum_{t=1}^{T} S(ves_{EQ,t},\ mes_{EQ,t}), \quad (5)$$

where $S(ves_{EQ,t},\ mes_{EQ,t})$ is defined as

$$S(ves_{EQ,t},\ mes_{EQ,t}) = \begin{cases} 4, \textit{ if the emotional quadrant and temporal phase are same} \\ 3, \textit{ if only the emotional quadrant is same} \\ 2, \textit{ if only the temporal phase is same} \\ 1, \textit{ otherwise} \end{cases}, \quad (6)$$

based on whether video and music are in the same emotional quadrant and with the same emotional temporal phase at time t. A ranked list of music (or video) is generated in descending order of $S(VES_{EQ},\ MES_{EQ})$ (or $S(MES_{EQ},\ VES_{EQ})$) over all testing pairs, and the top one is regarded as the best recommendation for the queried video (or music) to generate the MV.

4. EXPERIMENTS

To evaluate the effectiveness of the proposed EMV-matchmaker framework, we performed experiments on a set of OMVs downloaded from YouTube according to the links provided in the DEAP database [12]. 65 complete OMVs were collected, with a total length of about 5 hours. Each OMV was assigned one (out of three) emotional quadrant based on the VA annotations provided in the DEAP database. The two emotional quadrants in the low arousal space were merged into one, as shown in Figure 2, since emotions mapped into the lower arousal space are difficult to differentiate [13]. For music, we used MIRToolbox to extract four types of frame-based acoustic features, namely dynamic, spectral, timbre, and tonal features [14,15]. For video, the frame-based color themes and motion intensities were extracted as the visual features [16,17].

In the experiments, two tasks were conducted: video retrieving music (i.e., the "V2M" task) and music retrieving video (i.e., the "M2V" task). We compared the proposed EMV-matchmaker framework with the state-of-the-art AVEG framework (k=32, iteration=12) [6,14,18], with 5-fold cross-validation. For the "V2M" task, the 65 videos were divided into 5 disjoint subsets. For each fold, each video of the testing subset was used in turn to search for the best matched music, and the videos in the

Table 1. Average ranking accuracy of the EMV-matchmaker and AVEG frameworks.

The "V2M" Task (Video Retrieving Music)		
	AVEG [6]	**EMV-matchmaker**
complete dataset	0.6401	0.6706
outliers removed	0.5434	0.7369
The "M2V" Task (Music Retrieving Video)		
	AVEG [6]	**EMV-matchmaker**
complete dataset	0.6797	0.5986
outliers removed	0.6253	0.6464

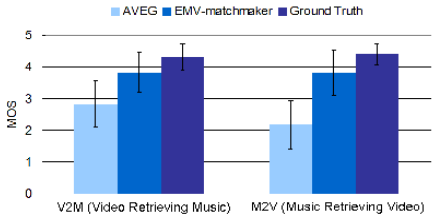

Figure 5. Results of the subjective MOS test.

remaining 4 subsets were used for training the visual ETCM to be used to generate the predicted emotional temporal phase sequence for the queried video. All the music tracks of the 65 OMVs were considered candidates and the one extracted from the test video was regarded as the ground truth. The ranking accuracy [5] defined as

$$Ranking\ Accuracy = 1 - \frac{rank(g) - 1}{|C| + 1}, \tag{7}$$

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 music clips in the candidate set. We reported the average ranking accuracy over all folds. The "M2V" task adopts the same experimental setup as the "V2M" task.

As shown in the rows marked by "complete dataset" in Table 1, the proposed EMV-matchmaker framework slightly outperformed the AVEG framework in the "V2M" task, but performed worse in the "M2V" task, although it is expected that EMV-matchmaker should outperform AVEG in both tasks as EMV-matchmaker matches video and music based on the detailed emotional temporal phase sequences while AVEG matches them simply based on average emotional representations. One possible reason could be that the dataset contains several MVs that are composed from live concerts and singing contests, and the emotional expressions of the video parts of these MVs are actually not as matched to those of the music parts as they are in the remaining OMVs. There were 21 such MVs. Therefore, we also evaluated the retrieval performance only on the remaining 44 queried video (or music) clips. The results are shown in the rows marked by "outliers removed" in Table 1. We can see that EMV-matchmaker outperformed AVEG in both the "V2M" and "M2V" tasks, and its performance was improved a lot. Interestingly, the performance of AVEG degraded. We believe that it is also because AVEG did not consider the temporal course of emotional expression.

Subjective evaluation[1] in terms of 5-point mean opinion score (MOS) was conducted on 6 "V2M" MV sets generated by retrieving music for video and 6 "M2V" MV sets. Each "V2M" (or "M2V") MV set contains the original official MV (ground truth) and the MVs generated by EMV-matchmaker and AVEG. Each of the 36 MVs (including automatically generated MVs and original official MVs) was evaluated by sixteen subjects. Note that the MVs were randomly selected from the results of the "complete dataset" experiments, but none were outliers. The average MOS over all MVs and subjects is shown in Figure 5. It is clear that EMV-matchmaker achieved better MOS performance than AVEG. The results reveal that considering the temporal course of emotional expression in video and music matching can generate more attractive MVs to enhance subjects' viewing and listening experiences. The results also show that the MOS of the MVs generated by EMV-matchmaker is quite close to that of the ground truth MVs. In fact, we did find that some MVs generated by EMV-matchmaker are satisfactory.

[1] MOS results for individual MVs and a system demonstration video are available at https://sites.google.com/site/emvmatchmaker/home?pli=1.

5. CONCLUSIONS AND FUTURE WORK

We have presented a novel content-based emotion-oriented music video generation system called EMV-matchmaker that utilizes the emotional temporal phase sequence of the multimedia content as a bridge to connect music and video. The results of both subjective and objective evaluations have demonstrated that the proposed EMV-matchmaker framework outperforms the state-of-the-art AVEG framework, and can offer a satisfactory generated music video to enhance human viewing and listening experience. Personalization techniques can be applied to further address the subjectivity issue of automatic MV generation, which is important and will be studied in our future work.

6. ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and Technology of Taiwan under Grant: NSC 102-2221-E-001-008-MY3. The first author would like to thank his best friend, Chao-Wei Chen, for his suggestion and encouragement.

7. REFERENCES

[1] P. N. Juslin and D. Västfjäll. Emotional responses to music: the need to consider underlying mechanisms. *Behav Brain Sci.*, 2008.
[2] D. A. Shamma et al. MusicStory: a personalized music video creator. In *ACM MM*, 2005.
[3] O. Gillet et al. On the correlation of automatic audio and visual segmentations of music videos. *IEEE TCSVT*, 2007.
[4] X. Wu et al. Automatic music video generation: cross matching of music and image. In *ACM MM*, 2012.
[5] F. F. Kuo et al. Background music recommendation for video based on multimodal latent semantic analysis. In *ICME*, 2013.
[6] J. C. Wang et al. The acousticvisual emotion Gaussians model for automatic generation of music video. In *ACM MM*, 2012.
[7] R. R. Shah et al. ADVISOR–personalized video soundtrack recommendation by late fusion with heuristic rankings. In *ACM MM*, 2014.
[8] R. E. Thayer. *The Biopsychology of Mood and Arousal*. New York: Oxford Univ. Press, 1989.
[9] P. Ekman. *Handbook of Cognition and Emotion*. Wiley, 1999.
[10] J. C. Lin et al. Emotion recognition of conversational affective speech using temporal course modeling. In *INTERSPEECH*, 2013.
[11] C. H. Wu et al. Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course. *IEEE TMM*, 2013.
[12] S. Koelstra, et al. DEAP: a database for emotion analysis using physiological signals. *IEEE TAC*, 2012.
[13] M. Soleymani et al. A Bayesian framework for video affective representation. In *ACII*, 2009.
[14] J. C. Wang et al. The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In *ACM MM*, 2012.
[15] O. Lartillot and P. Toiviainen. A Matlab toolbox for musical feature extraction from audio. In *DAFx*, 2007.
[16] X. Wang et al. Affective image adjustment with a single word. *Vis. Comput.*, 2013.
[17] H. W. Chen et al. Action movies segmentation and summarization based on tempo analysis. In *MIR*, 2004.
[18] Tool of the AEG model: http://slam.iis.sinica.edu.tw/demo/aeg/