The Acousticvisual Emotion Gaussians Model for Automatic Generation of Music Video*

Ju-Chiang Wang[1,2], Yi-Hsuan Yang[1], I-Hong Jhuo[1,2], Yen-Yu Lin[1] and Hsin-Min Wang[1]
[1] Academia Sinica, Taiwan; [2] National Taiwan University, Taiwan
{asriver, yang, ihjhuo, yylin, whm}@iis.sinica.edu.tw

ABSTRACT

This paper presents a novel content-based system that utilizes the perceived emotion of multimedia content as a bridge to connect music and video. Specifically, we propose a novel machine learning framework, called Acousticvisual Emotion Guassians (AVEG), to jointly learn the tripartite relationship among music, video, and emotion from an emotion-annotated corpus of music videos. For a music piece (or a video sequence), the AVEG model is applied to predict its emotion distribution in a stochastic emotion space from the corresponding low-level acoustic (resp. visual) features. Finally, music and video are matched by measuring the similarity between the two corresponding emotion distributions, based on a distance measure such as KL divergence.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]

Keywords

Computational emotion model, cross-modal media retrieval.

1. INTRODUCTION

Recent years have witnessed a tremendous growth of online video sharing on websites such as Youtube and Nico Nico Douga.[1] Such services have drastically changed the way multimedia content is created, distributed, and accessed. Everyone can easily create a video sequence with a consumer camcorder and broadcast it over the Internet.

To enhance the entertaining and aesthetic qualities of the video sequences, it is usually useful to accompany a video sequence with a piece of music that goes well together. For example, exciting music might be good company for sports video. Due to the absence of an effective automatic tool, this composition task is usually performed manually, which is extremely labor extensive when one needs to select from thousands of soundtracks for a given video sequence.

In response to this demand, machine-aided automatic music video composition has been studied in the last decade [1–3]. However, the performance of existing systems is usually limited, because most of the time only the relationship between low-level acoustic features (such as characteristics of the audio spectrogram) and visual features (such as color and texture) is considered. It is difficult to establish a direct relationship between the two modalities from low-level features. Moreover, there is a so-called semantic gap between low-level signal features and high-level human perception.

*This work was supported by the National Science Council of Taiwan under Grants: NSC-101-2221-E-001-019, NSC-101-2631-H-001-007, and NSC-100-2218-E-001-009.

[1] `http://www.youtube.com/`, `http://www.nicovideo.jp/`

MM'12, October 29–November 2, 2012, Nara, Japan.
ACM 978-1-4503-1089-5/12/10.

Motivated by the recent development in affective computing of multimedia signals [4–7], we propose to select a music piece that is in tune with the given video sequence with respect to the affective content. A music-accompanied video composed in this way is attractive, as the perception of emotion naturally occurs in video watching.

Specifically, we propose a novel machine learning model, called Acousticvisual Emotion Guassians (AVEG), to jointly learn the tripartite relationship among music, video, and emotion from an emotion-annotated corpus of music videos (MVs). Cross-modal factor analysis (CFA) [8] is employed to reduce the gap between the low-level acoustic and visual features in an unsupervised manner. Then, the AVEG model utilizes the features processed by CFA to make emotion prediction in an emotion space for music and video, respectively. Finally, the matching between music and video can be done in the emotion space via measuring the KL divergence between their corresponding predicted emotions.

2. EMOTION MODEL AND CORPUS

To identify the internal human representations of emotion, psychologists have applied factor analysis techniques such as multidimensional scaling to the emotion ratings of music stimuli. Although differ in names, existing studies give very similar interpretations of the resulting fundamental factors, most of which correspond to *valence* (positive/negative affective states), *activation* (or arousal; energy level), and *potency* (or dominance; a sense of control or freedom to act) [9]. The 3D emotion space is referred to as 3DES hereafter.

For reproducibility, the DEAP dataset [10] is utilized. It contains 120 pieces of 1-minute MV segments collected from Youtube. Each segment was on average annotated by 14-16 volunteers, who were asked to annotate valence, activation, and potency on a discrete 9-point scale online [10].

3. SYSTEM AND METHODOLOGY

3.1 Feature Extraction and CFA

We extract segment-level (with dynamic length) visual and acoustic features. The visual features are extracted via densely sampled trajectories in spatio-temporal volumes [11], and each segment for a trajectory is characterized by four different descriptors: *motion boundary histogram*, *histogram of oriented gradient*, *histogram of optical flow*, and *trajectory shape*. The acoustic features we adopt include those related to *timbre* (Mel-scale frequency cepstral coefficients), *pitch* (chroma and pitch features), sound *intensity* (root mean square energy), and *rhythm* (fluctuation pattern and tempo) [4, 7]. For an MV, each segment for extracting acoustic features is defined by the sampled trajectories from video, and the frame-based acoustic features in a segment are summarized by their mean and standard deviation.

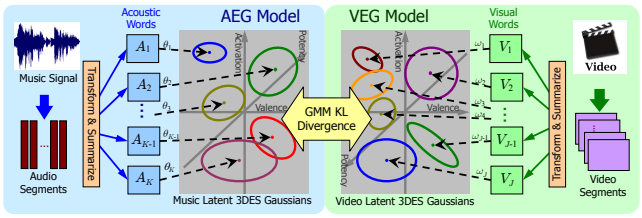

Figure 1: Illustration of the AVEG framework.

For better performance, we apply cross-modal factor analysis (CFA) [8] to align the acoustic and visual feature spaces and find the principle projections for them, respectively. Suppose that a pair of segment-level acoustic and visual features are represented as $\mathbf{x} \in \mathbb{R}^{d_a}$ and $\mathbf{y} \in \mathbb{R}^{d_v}$, two transformation matrices, $\mathbf{A} \in \mathbb{R}^{p \times d_a}$ and $\mathbf{B} \in \mathbb{R}^{q \times d_v}$, where $p \leq d_a$ and $q \leq d_v$ can be different and set empirically according to their original dimensions, are respectively learned via CFA from a universal set, $\mathcal{F}$, consisting of 10K segments randomly selected from the DEAP corpus. It follows that the segment-level feature spaces for the audio and video modalities can be respectively derived by $\hat{\mathbf{x}} \leftarrow \mathbf{A}\mathbf{x}$ and $\hat{\mathbf{y}} \leftarrow \mathbf{B}\mathbf{y}$.

3.2 Acousticvisual Emotion Gaussians

As shown in Figure 1, the AVEG framework contains two models, namely acoustic emotion Gaussians (AEG) and visual emotion Gaussians (VEG), which are conceptually the same except for difference in the media modality that is taken into account (audio and video, respectively). Note that due to the independence between AEG and VEG, these two models can be learned separately using different emotion annotated datasets, as long as the underlying emotion model is the same (e.g., 3DES). Below we briefly introduce the basic idea of the AEG model, since that for the VEG model is similar. Readers are referred to [12] for details.

We aggregate a set of segment-level $\hat{\mathbf{x}}$s of a music clip into a clip-level probabilistic vector by implementing the acoustic words $\{A_k\}_{k=1}^{K}$, as shown in Figure 1, with an acoustic Gaussian mixture model (GMM) pre-learned on $\mathcal{F}$. Accordingly, each training clip s_n in DEAP $\mathcal{D}$ is represented as a vector $\boldsymbol{\theta}_n$, whose k-th component is a posterior probability θ_{nk} corresponding to the k-th Gaussian A_k in the acoustic GMM. Let's denote the emotion annotations of s_n as $\{\mathbf{e}_{nm}\}_{m=1}^{U_n} \in \mathbb{R}^3$, where $\mathbf{e}_{nm}$ is given by the m-th subject of s_n. Then, given $\mathcal{D}$, all annotations, $\mathbf{E} \equiv \{\mathbf{e}_{nm}\}_{\forall n,m}$, can be generated from a weighted 3DES GMM with $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$,

$$p(\mathbf{E}|\mathcal{D}) = \prod_n \prod_m \sum_k \theta_{nk} \mathcal{N}(\mathbf{e}_{nm}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \qquad (1)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean vector and covariance matrix of the k-th latent 3DES Gaussian (cf. Figure 1), which can be learned using the EM-based algorithm [12].

Due to the parametric and probabilistic nature underlying AVEG, personalization can be achieved by adapting the 3DES GMMs in a dynamic and efficient manner when a small number of personal annotations are available [13].

3.3 Music and Video Matching

Suppose we have learned the acoustic and visual 3DES GMMs, denoted by $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ and $\{\boldsymbol{\eta}_j, \boldsymbol{\Lambda}_j\}_{j=1}^{J}$, respectively, where J can be unequal to K, the predicted 3DES distributions for a music clip s and video v respectively are

$$p(\mathbf{e}|s) = \sum_{k=1}^{K} \theta_k \mathcal{N}(\mathbf{e}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \qquad (2)$$

$$p(\mathbf{e}|v) = \sum_{j=1}^{J} \omega_j \mathcal{N}(\mathbf{e}|\boldsymbol{\eta}_j, \boldsymbol{\Lambda}_j)\,, \qquad (3)$$

where $\{\theta_k\}_{k=1}^{K}$ and $\{\omega_j\}_{j=1}^{J}$ are the feature posteriors of s and v derived from their corresponding feature words and segment-level feature vectors ($\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$), respectively.

As shown in Figure 1, since a music clip and a video sequence can be mapped into the same emotion space (cf. Eqs. 2 and 3), the relevance between them can be measured according to the KL divergence D_{KL} between their resulting 3DES GMMs. For better efficiency, we apply the variational based method [14] to compute the approximated lower bound of D_{KL} between $p(\mathbf{e}|s)$ and $p(\mathbf{e}|v)$. From Eqs. 2 and 3, the predicted emotions of music and video are estimated by the weighted 3DES GMMs using $\{\theta_k\}_{k=1}^{K}$ and $\{\omega_j\}_{j=1}^{J}$ as weights, respectively, and $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ and $\{\boldsymbol{\eta}_j, \boldsymbol{\Lambda}_j\}_{j=1}^{J}$ are fixed. We can therefore lessen the computational cost by computing the componentwise basis KL divergences between the two non-weighted 3DES GMMs beforehand in an off-line manner. As $\{\omega_j\}_{j=1}^{J}$ of a video query v is given, we can efficiently compute $D_{\mathrm{KL}}(v\|s)$ between v and each music piece s in the database. Likewise, for a music query s, $D_{\mathrm{KL}}(s\|v)$ can be used to retrieve videos. The complexity of the AVEG model and the matching procedure only depend on K and J. For mobile devices, one can select K and J for balancing the trade-off between accuracy and efficiency.

4. CONCLUSION

We have presented a novel system that utilizes the novel AVEG framework, which jointly learns the tripartite relationship among music, video, and emotion from an emotion-annotated corpus of MVs, to bridge music and video in a higher level 3DES. We will perform more subjective evaluations on the automatically generated music videos.

5. REFERENCES

[1] Shamma, A. et al. 2005. MusicStory: A personalized music video creator. *Proc. ACM MM*, pp. 563–566.

[2] Gillet, O. et al. 2007. On the correlation of automatic audio and visual segmentations of music videos. *IEEE TCSVT*, 17, 3, pp. 347–355.

[3] Yoon, J.-C., et al. 2009. Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools and Application*, 41, 2, pp. 197–214.

[4] Yang, Y.-H. and Chen, H. H. 2011. *Music Emotion Recognition*, Cambridge: CRC Press.

[5] Hanjalic, A. and Xu, L.-Q. 2005. Affective video content representation and modeling. *IEEE TMM*, 7, 1, pp. 143–154.

[6] Zhang, S., et. al. 2009. Utilizing affective analysis for efficient movie browsing. *Proc. ICIP*, pp. 1853–1856.

[7] Benini, S., et al. 2011. A connotative space for supporting movie affective recommendation. *IEEE TMM*, 13, 6, pp. 1356–1370.

[8] Li, D., et al. 2003. Multimedia content processing through cross-modal association. *Proc. ACM MM*, pp. 604–611.

[9] Fontaine, J. R., et al. 2007. The world of emotion is not two-dimensional. *Psychological Science*, 18, 2, pp. 1050–1057.

[10] Koelstra, S. et al. 2012. DEAP: A database for emotion analysis using physiological signals. *IEEE Trans. Affective Computing*, 3, 1, pp. 18–31.

[11] Wang, H., et al. 2011. Action recognition by dense trajectories. *Proc. CVPR*, pp. 3169–3176.

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

[13] Wang, J.-C, Yang, Y.-H., Wang, H.-M., and Jeng, S.-K. 2012. Personalized music emotion recognition via model adaptation. *Proc. APSIPA ASC.*

[14] Hershey, J. R. and Olsen, P. A. 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. *Proc. ICASSP*, pp. 317–320.