DEMV-MATCHMAKER: EMOTIONAL TEMPORAL COURSE REPRESENTATION AND DEEP SIMILARITY MATCHING FOR AUTOMATIC MUSIC VIDEO GENERATION

Jen-Chun Lin, Wen-Li Wei, and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taipei, Taiwan

{jenchunlin, lilijinjin}@gmail.com, whm@iis.sinica.edu.tw

ABSTRACT

This paper presents a deep similarity matching-based emotion-oriented music video (MV) generation system, called DEMV-matchmaker, which utilizes an emotion-oriented deep similarity matching (EDSM) metric 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. An emotional temporal structure preserved histogram (ETPH) representation is proposed to keep the recognized emotional temporal phase sequence information for EDSM metric construction. A deep neural network (DNN) is then applied to learn an EDSM metric based on the ETPHs for the given positive (official) and negative (artificial) MV examples. For MV generation, the EDSM metric is applied to measure the similarity between ETPHs of video and music. The results of objective and subjective experiments demonstrate that DEMV-matchmaker performs well and can generate appealing music videos that can enhance the viewing and listening experience.

Index Terms— Automatic music video generation, deep similarity learning, cross-modal media retrieval

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 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 movies and television programs. 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.

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 low-level acoustic (or visual) features and 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 an 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 low-level features (i.e., acoustic or visual features) and 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. Thus, we recently proposed an EMV-matchmaker framework [8], which consists of an emotional temporal course model (ETCM) to model the temporal structure of emotional expression and a stream matching method to measure the similarity between the recognized emotional temporal phase sequences of music and video, for music video generation. Although the EMV-matchmaker framework has been proved to outperform the state-of-the-art acoustic-visual emotion Gaussians (AVEG) framework [6], the fixed rigid similarity metric (i.e., string matching) used in EMV-matchmaker may not be always optimal. It cannot accommodate to the incorrect recognition of emotional temporal phase sequence.

To handle the aforementioned problem, inspired by the recent advance in deep learning [9–13] and distance metric learning (DML) [14–19] techniques, we first attempt to apply a deep neural network (DNN) to learn a flexible nonlinear similarity matching metric to alleviate the effect of recognition errors in an emotional temporal phase sequence for MV generation. DML has been extensively studied in both machine learning and multimedia communities. The crux of similarity search lies in two key components: (i) an effective feature representation and (ii) a proper similarity matching function over the feature space. Existing DML methods can be grouped into different categories according to varied learning settings and methodologies. For example, most DML studies in multimedia mainly learn metrics from various types of side-information, including class labels or binary

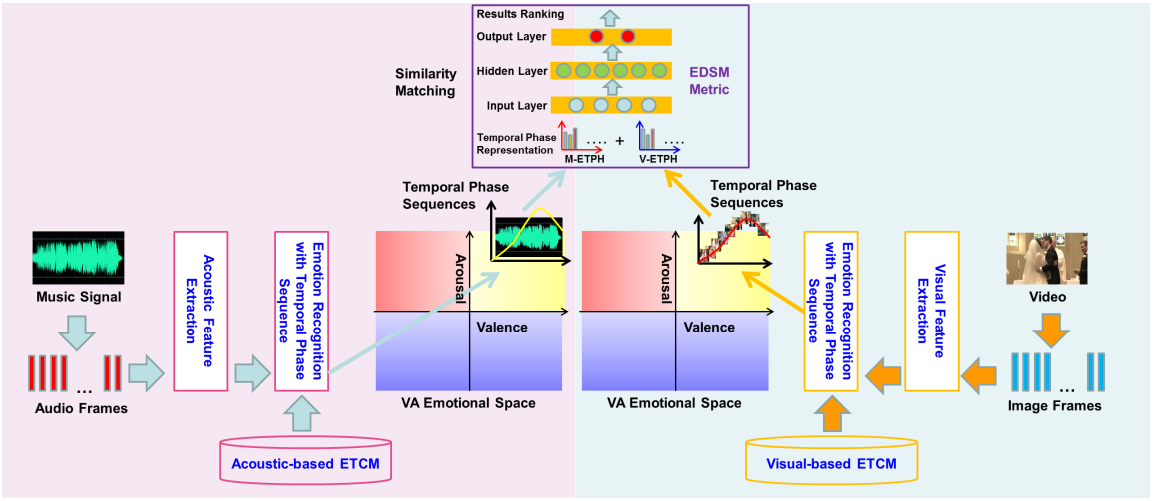

Figure 1. Illustration of the DEMV-matchmaker framework.

similar/dissimilar pairwise labels [15]. However, most learning methods aim to learn a linear distance metric in the form of Mahalanobis distance, which can be viewed as learning a linear projection to map the input feature space into another feature space. The linearity assumption inevitably limits the capacity of similarity measure for complex patterns. To tackle such challenges, inspired by the recent advance in deep learning techniques [9–13], researchers have begun to adopt a DNN to learn a nonlinear transformation function for similarity measure in many multimedia applications. For example, Hu et al. [17] proposed a discriminative deep metric learning method, which trains a DNN to learn a set of hierarchical nonlinear transformations to project face pairs into the same feature subspace for face verification in the wild. Srivastava et al. [16] employed a deep belief network to learn a generative model of the joint space of image and text inputs for cross-media information retrieval. Wu et al. [18] proposed an online multimodal deep similarity learning framework to learn a nonlinear transformation function for each feature modality and find an optimal combination of multiple modalities for image retrieval.

In this paper, as shown in Figure 1, a deep similarity matching-based emotion-oriented music video generation system, called DEMV-matchmaker, is proposed, as an extended framework of the previous EMV-matchmaker. For a music (or video) clip, an acoustic (or visual) emotional temporal course model (ETCM) is firstly applied to recognize its emotional temporal phase sequence in an emotional quadrant in the valence-arousal (VA) emotional space [20] from its low-level acoustic (or visual) features. For MV generation, an emotion-oriented deep similarity matching (EDSM) metric is then applied to match music and video clips based on whether they are with similar emotional temporal structure preserved histogram (ETPH) representations. The acoustic and visual ETCMs can be learned from an emotion-annotated MV corpus. The EDSM metric can be learned for the given positive (official) and negative (artificial) MV examples. To the best of our knowledge, this is the first attempt to consider the deep similarity learning technique for emotion-oriented MV generation.

2. METHODOLOGY

In the proposed DEMV-matchmaker system, as shown in Figure 1, 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. An EDSM metric is then used to match music and video clips based on their ETPHs for MV generation.

2.1. Emotional Temporal Phase Sequence Recognition and Representation

The psychologist Ekman's research [21] 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 recognize the temporal course of emotional expression of a MV (including music and video contents), the ETCM developed in our previous work [8] is adopted. As shown in Figure 1, the DEMV-matchmaker framework contains one acoustic ETCM and one visual ETCM for modeling music and video contents, respectively.

2.1.1. 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[1] $EQ \in \{EQ_1, EQ_2, EQ_3\}$ in the VA space shown in Figure 1, 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,

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

[1] The two emotional quadrants in the low arousal space were merged into one, as shown in Figure 1, since emotions mapped into the lower arousal space are difficult to differentiate [22].

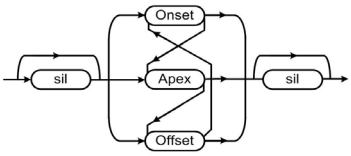

Figure 2. Recognition network based on the predefined grammar for characterizing an emotional quadrant expressed in a music or video clip.

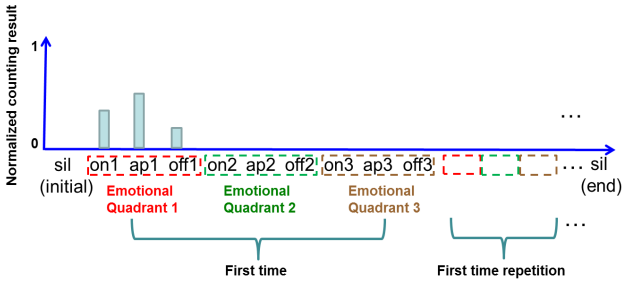

Figure 3. Illustration of the construction of the ETPH representation, where on1, ap1, and off1 represent the onset, apex, and offset phases of emotional quadrant 1.

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 HMMs 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 2; $P(O)$ 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

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

In ETCM training, for each emotional quadrant, we trained a set of HMMs (i.e., the acoustic (or visual) emotional sub-state HMMs, including the onset HMM, the apex HMM, and the offset HMM) from a set of official music videos (OMVs) that is annotated with the emotional temporal phase sequences, 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. In addition, we also trained the sil HMMs (cf. Figure 2) 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 repetition of emotional temporal phases in an OMV, an emotional temporal course grammar, as shown in Figure 2, was used to guide the recognition process by referring to the emotional temporal phases. All the temporal phase transition probabilities in the grammar were assumed uniformly distributed in this study.

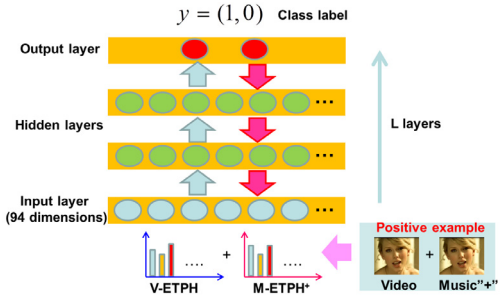

Figure 4. Illustration of the EDSM metric learning, where blue and red arrows represent the forward and back propagation procedures, respectively.

2.1.2. Emotional Temporal Phase Sequence Representation

Since a neural network cannot model a variable-length sequential input sequence, we convert an emotional temporal phase sequence into a fixed-dimensional emotional temporal structure preserved histogram (ETPH) vector for EDSM metric learning. Figure 3 illustrates the construction of the ETPH vector, where the horizontal and vertical axes represent the ordered emotional temporal phase index and the corresponding normalized count in the recognized emotional temporal phase sequence, respectively. This example is recognized to be in quadrant 1; thus, the counts for the other quadrants are set to 0. 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, we extract the first 5 repetitions in ETPH construction, even if a video or music clip contains more repetitions. For those video or music clips with fewer repetitions, the counts are set to 0 for the non-existing repetitions. In this way, the recognized emotional temporal phase sequence of a video (or music) clip is represented as a 47-dimensional V-ETPH (or M-ETPH) vector.

2.2. Emotion-oriented Deep Similarity Learning and Matching for MV Generation

In this study, we regard the similarity learning problem as a classification similarity learning issue [19]. The goal is to learn a classifier (i.e., EDSM metric) that can decide whether a pair of V-ETPH and M-ETPH is similar. In EDSM metric learning, as shown in Figure 4, a DNN is adopted to learn the classifier based on a set of positive training examples x^+=(V-ETPH, M-ETPH$^+$) and negative training examples x^-=(V-ETPH, M-ETPH$^-$) with binary class labels y^+=(1,0) and y^-=(0,1), respectively. A positive training example is directly extracted from the V-ETPH and M-ETPH of an OMV; while a negative training example is artificially constructed from the V-ETPH of an OMV and the M-ETPH of another OMV in a different emotional quadrant.

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

$$x^{(l+1)} = f^{(l)}(W^{(l)} x^{(l)} + b^{(l)}) \quad (5)$$

where $x^{(l)}$ and $x^{(l+1)}$ 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 a non-linear activation function, which is a *sigmoid* function for l=1 to L-2, and a *softmax* function for l=L-1.

Given the class label y, we use the *softmax* regression [23] as the loss function in the output layer as

$$\ell(x, y) = KL(x^{(L)}, y) \quad (6)$$

where $KL(.)$ is the KL-divergence function. 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 side-information (i.e., positive or negative class label) is considered for DNN to learn a nonlinear similarity matching metric, the constructed DNN classifier (i.e., EDSM metric) is expected to alleviate the effect of recognition errors in an emotional temporal phase sequence.

In the MV generation phase, given a queried video clip, the goal is to find a ranked list of music clips for the query. Specifically, the queried video clip is paired with each music clip from the target music database to form a testing pair. The visual and acoustic ETCMs as well as the ETPH representation method are applied to obtain the corresponding V-ETPH and M-ETPH. The EDSM metric is then applied to measure the similarity between the V-ETPH and M-ETPH. Finally, all the music clips are ranked in descending order of scores obtained from the first output node of the EDSM metric, and the top one is regarded as the best recommendation for the queried video to generate the MV.

Since every matching metric has its own advantage and disadvantage, we may further combine the ranking result of the proposed DEMV-matchmaker with that of the string matching-based EMV-matchmaker [8] (we name the combined system "DEMV-matchmaker$_{COM}$") as

$$R_{COM}(V, M)=R_{EMV}(V, M)+R_{DEMV}(V, M) \quad (7)$$

where $R_{COM}(V, M)$ represents the combined rank for an arbitrary V-M pair; R_{EMV} is the rank given by EMV-matchmaker; and R_{DEMV} is the rank given by DEMV-matchmaker. Finally, all the music clips are ranked in ascending order of R_{COM}, and the top one is regarded as the best recommendation for the queried video.

3. EXPERIMENTS

To evaluate the effectiveness of the proposed DEMV-matchmaker 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 [24] were used to train the acoustic ETCM, the visual ETCM, and the EDSM metric. Each OMV was assigned one (out of three) emotional quadrant based on the VA annotations provided in the DEAP database. The emotional temporal phases of each OMV were annotated according to the repetitions of verse-chorus sections by referring to the lyrics [8]. 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 [25,26]. For video, the frame-based color themes and motion intensities were extracted as the visual features [27,28]. For training the EDSM metric, we used a DNN with 4 hidden layers with 100, 100, 80, and 10 neurons, respectively. We applied random initialization for the weights, a constant learning rate of 0.001, and the L2 weight decay regularization to avoid over-fitting. The size of mini-batch for the stochastic gradient descent algorithm was set to 10.

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 [5] defined as

$$Ranking\ Accuracy = 1 - \frac{rank(g) - 1}{|C| + 1}, \quad (8)$$

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 ($|C|$=200 in this study). We reported the average ranking accuracy over the testing set.

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

The Video Retrieving Music (V2M) Task		
EMV-matchmaker [8]	**DEMV-matchmaker**	**DEMV-matchmaker$_{COM}$**
0.6057	0.6414	0.6519

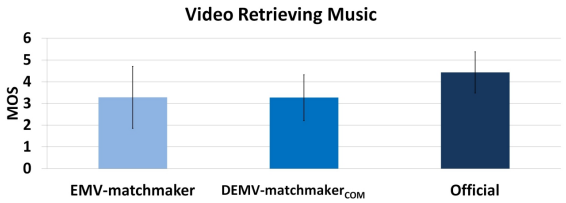

Figure 5. Results of subjective evaluation.

The results in Table 1 demonstrate that the new DEMV-matchmaker framework outperforms the previous EMV-matchmaker framework. We believe that it is because a fixed rigid similarity metric (i.e., string matching) used in EMV-matchmaker cannot accommodate to the incorrect recognition of emotional temporal phase sequence. A learning-based nonlinear similarity matching metric (i.e., EDSM metric) used in DEMV-matchmaker indeed alleviates to some extent the effect of recognition errors. DEMV-matchmaker$_{COM}$ further improves the performance. Overall, DEMV-matchmaker$_{COM}$ pushed ahead the rank of ground truth music by approximately 10 (i.e., the average ranking accuracy was improved from 0.6057 to 0.6519), compared to EMV-matchmaker.

Subjective evaluation[2] in terms of 5-point mean opinion score (MOS) was conducted on 5 MV sets. Each MV set contains the original official MV (ground truth) and the MVs generated by EMV-matchmaker and DEMV-matchmaker$_{COM}$. Each MV was evaluated by 14 subjects. The average MOS results in Figure 5 show that DEMV-matchmaker$_{COM}$ only very slightly outperforms EMV-matchmaker, and that the gap between an automatically generated MV and an official MV is quite small.

4. CONCLUSIONS AND FUTURE WORK

We have presented a deep similarity matching-based emotion-oriented music video (MV) generation system, called DEMV-matchmaker, which utilizes an emotion-oriented deep similarity matching (EDSM) metric as a bridge to connect music and video. The results of both subjective and objective evaluations have demonstrated that the proposed DEMV-matchmaker framework outperforms the state-of-the-art EMV-matchmaker framework, and can offer a satisfactory automatically generated music video to enhance human viewing and listening experience. Different DNN structures and objective functions can be applied to further address the similarity matching issue of automatic MV generation, which is important and will be studied in our future work.

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

5. REFERENCES

[1] P. N. Juslin and D. Västfjäll, "Emotional responses to music: the need to consider underlying mechanisms," *Behav Brain Sci.*, vol. 31, no. 5, pp. 559–621, 2008.

[2] D. A. Shamma, B. Pardo, and K. J. Hammond, "Musicstory: a personalized music video creator," In *ACM MM*, 2005.

[3] O. Gillet, S. Essid, and G. Richard, "On the correlation of automatic audio and visual segmentations of music videos," *IEEE TCSVT*, vol. 17, no. 3, pp. 347–355, 2007.

[4] X. Wu, B. Xu, Y. Qiao, and X. Tang, "Automatic music video generation: cross matching of music and image," In *ACM MM*, 2012.

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

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

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

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

[9] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[10] G. E. Hinton, S. Osindero, and Y.-W. The, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[11] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[12] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE TASLP*, vol. 20, no. 1, pp. 30–42, 2012.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In *NIPS*, 2012.

[14] K. Weinberger, J. Blitzer, and L. Saul, "Distance metric learning for large margin nearest neighbor classification," In *NIPS*, 2006.

[15] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *Journal of Machine Learning Research*, vol. 12, pp. 491−523, 2011.

[16] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," In *NIPS*, 2012.

[17] J. Hu, J. Lu, and Y. P. Tan, "Discriminative deep metric learning for face verification in the wild," In *CVPR*, 2014.

[18] P. Wu, S. C. H. Hoi, H. Xia, P. Zhao, D. Wang, and C. Miao, "Online multimodal deep similarity learning with application to image retrieval," In *ACM MM*, 2013.

[19] E. López-Iñesta, F. Grimaldo, and M. Arevalillo-Herráez, "Classification similarity learning using feature-based and distance-based representations: A comparative study," *An International Journal Applied Artificial Intelligence*, vol. 29, no. 5, pp. 445–458, 2015.

[20] R. E. Thayer. *The Biopsychology of Mood and Arousal*. New York: Oxford Univ. Press, 1989.

[21] P. Ekman. *Handbook of Cognition and Emotion*. Wiley, 1999.

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

[23] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao, "Cross-media hashing with neural networks," In *ACM MM*, 2014.

[24] S. Koelstra, et al., "DEAP: a database for emotion analysis using physiological signals," *IEEE TAC*, vol. 3, no. 1, pp. 18–31, 2012.

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

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

[27] X. Wang, J. Jia, and L. Cai, "Affective image adjustment with a single word," *Vis. Comput.*, vol. 29, no. 11, pp. 1121–1133, 2013.

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