Automatic Music Video Generation Based on Simultaneous Soundtrack Recommendation and Video Editing

Jen-Chun Lin[1], Wen-Li Wei[1], James Yang[2], Hsin-Min Wang[1], and Hong-Yuan Mark Liao[1]
[1]Academia Sinica, Taiwan; [2]National Chiao Tung University, Taiwan
{jenchunlin, lilijinjin, jamesyang00712}@gmail.com; {whm, liao}@iis.sinica.edu.tw

ABSTRACT

An automated process that can suggest a soundtrack to a user-generated video (UGV) and make the UGV a music-compliant professional-like video is challenging but desirable. To this end, this paper presents an automatic music video (MV) generation system that conducts soundtrack recommendation and video editing simultaneously. Given a long UGV, it is first divided into a sequence of fixed-length short (e.g., 2 seconds) segments, and then a multi-task deep neural network (MDNN) is applied to predict the pseudo acoustic (music) features (or called the pseudo song) from the visual (video) features of each video segment. In this way, the distance between any pair of video and music segments of same length can be computed in the music feature space. Second, the sequence of pseudo acoustic (music) features of the UGV and the sequence of the acoustic (music) features of each music track in the music collection are temporarily aligned by the dynamic time warping (DTW) algorithm with a pseudo-song-based deep similarity matching (PDSM) metric. Third, for each music track, the video editing module selects and concatenates the segments of the UGV based on the target and concatenation costs given by a pseudo-song-based deep concatenation cost (PDCC) metric according to the DTW-aligned result to generate a music-compliant professional-like video. Finally, all the generated MVs are ranked, and the best MV is recommended to the user. The MDNN for pseudo song prediction and the PDSM and PDCC metrics are trained by an annotated official music video (OMV) corpus. The results of objective and subjective experiments demonstrate that the proposed system 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

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123399

1 INTRODUCTION

With the prevalence of mobile devices, video is widely used to record memorable moments such as weddings, graduations, and birthday parties. Popular websites such as YouTube and Vimeo have further boosted the phenomenon as broadcasting becomes easy. However, in music videos (MVs), movies, and television programs, music and video are often accompanied to complement each other to enhance emotional resonance and viewing experiences. Without soundtracks, most user-generated videos (UGVs) might look boring. Therefore, accompanying a UGV with music to enhance the entertaining quality and emotional resonance is highly desirable. For example, a wedding video with a romantic soundtrack can enhance a sweet atmosphere. Nevertheless, selecting the right music for a video requires music professionals. With the rapid growth of music collections, the task becomes even more difficult. Furthermore, a UGV often consists of unedited long-running and redundant content. Manually editing through the staggering number of images of a UGV entails significant human labor. Therefore, an automated process, which can edit a long UGV into a music-compliant professional-like video, is preferable. Under these circumstances, we propose a fully automatic MV generation system, which is able to conduct soundtrack recommendation and video editing simultaneously.

Machine-aided MV composition has been studied over the past decade [1–8]. However, previous research addressed either video editing or soundtrack recommendation [1–8], and none of them handled both tasks simultaneously. In the early period, most research effort was devoted to video editing [1–3]. Given a music track specified by a user, the goal is to generate an MV by selecting and concatenating suitable video clips to best match the music track [1–3]. The performance of such systems is usually limited, because only the relation between the low-level acoustic and visual features is considered, while the higher-level semantics (e.g., emotion) that can really catch the feeling of a human being is not. Even the low-level visual features of the selected video clips match well with the low-level acoustic features of music, the mismatched semantic structures may still result in bad viewing and listening experiences. For example, it has been demonstrated in [7] that the nonsynchronous temporal courses of emotional expression between video and music will result in bad viewing experiences. Moreover, there is a so-called semantic gap between the low-level acoustic (or visual) features and the high-level human perception. Different from editing a video to fit a specified music track, soundtrack recommendation is to suggest a matched music track to the video. Motivated by the recent development in

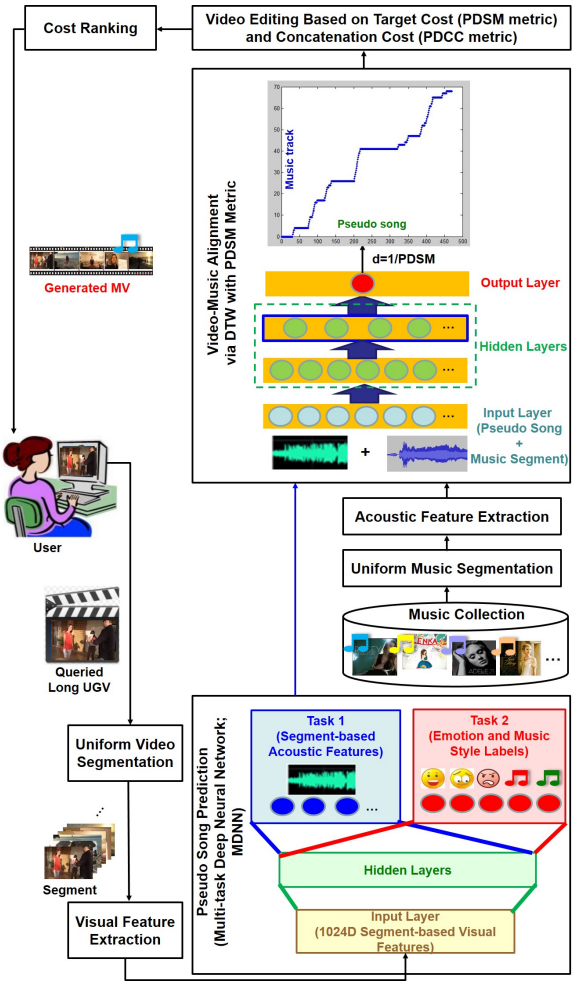

Figure 1: The proposed MV generation framework based on semantic-oriented pseudo song prediction, matching, and video editing.

affective computing of multimedia signals, most soundtrack recommendation systems map the low-level acoustic and visual features into an emotional space [5–8], and match these two modalities there. A music-accompanied video composed in this way could be attractive, as the perception of emotion naturally occurs during video watching and music listening. However, these systems [5–8] model the relation 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 relation between the music and video modalities in soundtrack recommendation may still result in bad viewing and listening experiences. In addition, the redundant content in a long unedited UGV will also dramatically ruin the viewing experience, even with a soundtrack. Such redundancy will inevitably cause a nonsynchronous emotional storytelling between the music and the UGV. In summary, an advanced MV generation system should address both video editing and soundtrack recommendation. The correlation among music, video, and semantic annotations such as emotion should also be actively explored and modeled.

Our idea to jointly handle the aforementioned problems is first 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 semantic-oriented pseudo song prediction, matching, and video editing, as shown in Figure 1. Given a long UGV, it is first divided into a sequence of fixed-length short (e.g., 2 seconds) segments. For each video segment, a multi-task deep neural network (MDNN) [12,13] is used to predict the pseudo acoustic (music) features (or called the pseudo song) from the visual (video) features of the video segment. The MDNN is trained by jointly learning the relation among acoustic (music) features, visual (video) features, and semantic annotations including emotion and music style labels from an annotated OMV corpus. Each music track in the music collection is also divided into a sequence of fixed-length short segments. Soundtrack recommendation and video editing are conducted simultaneously by aligning the UGV and a candidate music track in the acoustic (music) feature space using a sequence alignment technique. As shown in Figure 1, a dynamic time warping (DTW) algorithm with a pseudo-song-based deep similarity matching (PDSM) metric is applied to align the UGV and the candidate music track by evaluating the similarity between the pseudo acoustic features of the segments of the UGV and the acoustic features of the segments of the candidate music track. The PDSM metric is realized by a deep neural network (DNN) trained on the positive (official), less-positive (artificial), and negative (artificial) MV examples. The video editing module based on the target and concatenation costs then selects and concatenates the segments of the UGV according to the DTW-alignment with the candidate music track to generate a music-compliant professional-like video. The pseudo-song-based deep concatenation cost (PDCC) metric for evaluating the concatenation cost is another DNN designed to learn the visual storytelling, which can judge whether the concatenation of two video segments conforms to a professional way. The target cost is given by the PDSM metric. Finally, the cost ranking module will rank all the generated MVs and recommend the best MV to the user. Under this framework, our system not only recommends the matched music track to the UGV but also edits the UGV into a music-compliant professional-like video. That is, the length and the content of the UGV are edited to fit the matched music track. In summary, the main contributions of this paper include:

1. We have implemented a complete automatic music video generation system that can automatically edit a long user-

generated video into a music-compliant professional-like video.

2. This is the first work to simultaneously address both soundtrack recommendation and video editing issues in automatic MV generation.
3. We explore the correlation among music, video, and semantic annotations in OMVs by using the MDNN.
4. We propose the PDSM metric to alleviate the impact of pseudo song prediction errors on the similarity measure between the pseudo song and the candidate music.
5. We propose the PDCC metric for video editing to select video segments to concatenate in a more professional way.

The remainder of this paper is organized as follows. Previous research on video editing and soundtrack recommendation is reviewed in Section 2. The methodology is described in detail 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 recent progress on video editing and soundtrack recommendation in automatic MV generation. In the field of video editing, Hua et al. [1] employed the correlation coefficient to measure the correlation between the tempo sequence (music patterns) and the motion intensity sequence (video scene series) to select the scenes of video that best matched the specified music for MV generation. Wang et al. [2] employed a dynamic programming algorithm to measure the similarity between video shot attributes (i.e., normalized duration and motion) and music beat attributes (i.e., normalized length and tempo) to find a set of video shots that best matched the music track. They further analyzed the music structure (e.g., verse and chorus) and considered the events in a broadcast sports video in the extended work [3]. For soundtrack recommendation, Kuo et al. [4] employed multi-modal latent semantic analysis to learn the co-occurrence of the low-level acoustic and visual features, such as Mel-frequency cepstral coefficients, loudness, spectral centroid, color, and motion. 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. Shah et al. [6] employed a support vector machine (SVM) to model the categorical emotion, including sweet, funny, and sad, from the acoustic, visual, and geographic features. 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. 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 our MV generation system, as shown in Figure 1, uniform segmentation is first applied to divide a UGV into a sequence of fixed-length video segments. For each video segment, an MDNN is used to predict the acoustic (music) features from the visual (video) features, and the process is called pseudo song prediction. A DTW algorithm with a PDSM metric is used to align the UGV with a music track based on the acoustic features. Finally, the video editing and cost ranking modules generate the MV.

3.1 Video Segmentation

Consider that a UGV usually contains tens of thousands of image frames, it would be more efficient to conduct pseudo song prediction, matching, and video editing at the segment level instead of the frame level. In addition, since the DTW algorithm for aligning a UGV with a candidate music track requires an equal measurement unit, both the UGV and the music track are uniformly segmented into a sequence of fixed-length (e.g., 2 seconds) segments. The segment-based visual (or acoustic) features are constructed by the statistics of the component frame-based visual (or acoustic) features.

3.2 Pseudo Song Prediction via a Multi-Task Deep Neural Network

Multi-task learning [14] 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, 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 an MDNN [12,13] to predict the acoustic (music) features from the visual (video) features by jointly learning the relation among the acoustic (music) features, visual (video) features, and semantic annotations including emotion and music style labels from an annotated OMV database.

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

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

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.

We use emotion (i.e., the VA emotional quadrant labels) and music style (i.e., the slow-rhythm and fast-rhythm labels) 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 and music style labels provide a semantic constraint for MDNN learning so as 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,

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

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 segment-based visual feature vector constructed from the component frame-based visual feature vectors, $d_i^{(1)}$ is the i-th element of the corresponding segment-based acoustic feature vector, $d_i^{(2)}$ is the i-th semantic label, $s_i^{(1)}$ is the i-th element of the predicted pseudo acoustic feature vector, while $s_i^{(2)}$ is the predicted value of the i-th semantic label.

3.3 Video-Music Alignment via DTW

We assume that a UGV is longer than a candidate music track. Therefore, a DTW algorithm is used to find the best alignment between them. For a short UGV, there is no need to perform video-music alignment and video editing, as will be explained later in Section 4.

Let $\mathbf{P}\equiv\{\mathbf{p}_t\}_{t=1}^{T}$ and $\mathbf{C}\equiv\{\mathbf{c}_m\}_{m=1}^{M}$ be, respectively, the pseudo acoustic feature vector sequence of the UGV and the acoustic feature vector sequence of the candidate music track, where T and M denote the length, $T>M$, and $\mathbf{p}_t$ and $\mathbf{c}_m$ represent the pseudo acoustic feature vector for the t-th segment of the UGV and the acoustic feature vector for the m-th segment of the candidate music track. To find the time alignment between $\mathbf{P}$ and $\mathbf{C}$, a $T\times M$ distance matrix $\mathbf{D}=\left[D(t,m)\right]_{T\times M}$ is constructed, where $D(t,m)$ is the distance between $\{\mathbf{p}_1,\mathbf{p}_2,...,\mathbf{p}_t\}$ and $\{\mathbf{c}_1,\mathbf{c}_2,...,\mathbf{c}_m\}$ computed by

$$D(t,m)=\begin{cases} d(t,m)+\min\begin{cases} D(t-1,m) \\ D(t-1,m-1)\end{cases} & \forall t>1,\forall m>1 \\ d(t,m) & t=1,m=1 \\ \infty & t=1,\forall m>1 \\ d(t,m)+D(t-1,m) & \forall t>1,m=1\end{cases}, \tag{3}$$

where $d(t,m)$ is the distance between $\mathbf{p}_t$ and $\mathbf{c}_m$.

Since the goal is to edit the UGV into a music-compliant professional-like video, we allow a single music segment to be aligned with consecutive video segments in the UGV, but do not allow a single video segment in the UGV to be aligned with consecutive music segments in the candidate music track. That is, for any given node (t,m), $\forall t>1$ and $\forall m>1$, in the path, the possible fan-in nodes are restricted to node $(t-1,m)$ and node $(t-1,m-1)$ (cf. (3)).

After the distance matrix $\mathbf{D}$ has been constructed, the distance between $\mathbf{P}$ and $\mathbf{C}$ is evaluated as $D(T,M)$. The time-aligned path between $\mathbf{P}$ and $\mathbf{C}$ can be obtained by back tracking, and then used in the following video editing and cost ranking modules.

3.4 Pseudo-song-based Deep Similarity Metric

Directly calculating the distance between $\mathbf{p}_t$ and $\mathbf{c}_m$ by a rigid metric (e.g., the absolute or Euclidean distance) may not work well here, because a rigid distance metric cannot accommodate the prediction errors in the pseudo acoustic feature vector. To this end, a PDSM metric is adopted to learn a flexible nonlinear similarity metric to alleviate the impact of the prediction errors on the similarity measure between $\mathbf{p}_t$ and $\mathbf{c}_m$. That is, $d(t,m)$ in (3) is computed as the reciprocal of the score of the PDSM metric

$$d(t,m)=1/PDSM(\mathbf{p}_t,\mathbf{c}_m). \tag{4}$$

In this study, we regard PDSM metric learning as a regression 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 a music segment 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 segment associated with a video segment of an OMV. A less-positive training example is constructed from the pseudo song of a video segment of an OMV and the music segment of another OMV in the same VA emotional quadrant. A negative training example is constructed from the pseudo song of a video segment of an OMV and the music segment of another OMV in a different VA emotional quadrant.

By denoting a training example v^{++}, v^{+} or v^{-} as v, we forward v layer-by-layer through the DNN to generate the representation of each layer, i.e., $v^{(1)},\ldots,v^{(L)}$. The l-th layer takes $v^{(l)}$ as input and transforms $v^{(l)}$ to the output $v^{(l+1)}$ as follows,

$$v^{(l+1)}=f^{(l)}(W^{(l)}v^{(l)}+b^{(l)}), \tag{5}$$

where $W^{(l)}$ 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, the loss function in the output layer is the sum of squared error (SSE),

$$\ell(v,y)=SSE(v^{(L)},y). \tag{6}$$

The loss of the output layer will be back propagated to fine-tune the parameters W and b through a back-propagation method.

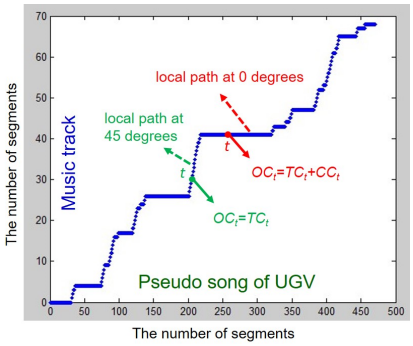

Figure 2: Illustration of the DTW-derived time-aligned path for a UGV and a candidate music track and the computation of the overall cost (OC) for a local path at 0 and 45 degrees.

Since the side-information (i.e., the positive, less-positive or negative label) is considered in learning a nonlinear similarity matching metric, the resulting DNN regression model (i.e., the PDSM metric) is expected to alleviate the impact of the pseudo-song prediction errors on the similarity measure. The difference between the proposed PDSM metric and the EDSM metric in [8] is that we regard similarity learning as a regression problem rather than a classification problem. By additionally considering the less-positive training examples, the learned PDSM metric should have a better generalization ability.

3.5 Video Editing Based on the Target and Concatenation Costs

After obtaining the time-aligned path for a UGV and a candidate music track, in the video editing stage, we first keep the segments of the UGV that correspond to the local paths at 45 degrees, as shown in Figure 2. Because such locations indicate the synchronization of the time series between the video segments and the music segments. As a result, the remaining issue of video editing is how to select, for a music segment corresponding to consecutive video segments (cf. the local paths at zero degrees in Figure 2), a suitable video segment from the consecutive video segments. Our idea to address this issue is inspired by the target and concatenation costs widely used in unit selection-based text-to-speech (TTS) synthesis [15,16], whose goal is to synthesize a speech utterance that pronounces a given sentence with fluent quality. The target cost is to estimate the perceptual difference between a target speech unit and a candidate speech unit, while the concatenation cost is to reflect a level of perceived discontinuity between consecutive speech units [15]. To edit the UGV into a music-compliant professional-like video, we apply the target and concatenation costs to select the suitable video segment from a local path at zero degrees. That is, the selected video segment should not only match the target music segment well (with a low target cost) but also keep the continuity (with a low concatenation cost) with the immediately preceding and succeeding video segments that are already selected (associated with the local path at 45 degrees).

Checking through the time-aligned path, the target cost (TC) for the t-th segment of the UGV is defined as $TC_t = d(t,m)$, the distance between its pseudo acoustic feature vector and the acoustic feature vector of the aligned m-th segment of the candidate music track. For the concatenation cost (CC), we use a pseudo-song-based deep concatenation cost (PDCC) metric to learn the visual storytelling literacy of professional video directors. Again, the learning of the PDCC metric is regarded as a regression problem. The goal is to learn a regression model (i.e., the PDCC metric) that can judge whether the concatenation of the pseudo acoustic features of two video segments conforms to a professional way. Similar to the PDSM metric learning, a DNN is adopted to learn a regression model based on a set of positive training examples c^{++}=(the pseudo songs of two consecutive video segments), less-positive training examples c^{+}=(the pseudo songs of two video segments separated by one segment), and negative training examples c^{-}=(the pseudo songs of two video segments separated by two segments) with labels u^{++}=3, u^{+}=2, and u^{-}=1, respectively. The increase of segment interval indicates the decrease of continuity. The training procedure of the PDCC metric is the same as that of the PDSM metric in Subsection 3.4. The resulting DNN regression model (i.e., the PDCC metric) is expected to have the ability to judge whether the concatenation of two video segments conforms to a professional way.

For a video segment t from a local path at zero degrees, the concatenation cost CC_t is calculated as

$$CC_t = \frac{1}{2}\left(1/PDCC(\mathbf{p}_{pcd}^{45^\circ}, \mathbf{p}_t^{0^\circ}) + 1/PDCC(\mathbf{p}_t^{0^\circ}, \mathbf{p}_{scd}^{45^\circ})\right), \quad (7)$$

where $\mathbf{p}_t^{0^\circ}$ denotes the pseudo acoustic feature vector of the t-th segment; $\mathbf{p}_{pcd}^{45^\circ}$ and $\mathbf{p}_{scd}^{45^\circ}$ denote the pseudo acoustic feature vectors of the preceding and succeeding segments from the local paths at 45 degrees. Then, the overall cost (OC) is calculated as

$$OC_t = TC_t + CC_t. \quad (8)$$

For each local path at zero degrees, the t^*-th segment that has the minimum OC will be included in the edited video. In this way, we can edit the UGV into the music-compliant professional-like video, i.e., both the length and the content of the UGV are edited to fit a specific candidate music track.

3.6 MV Generation via Cost Ranking

The final stage of the MV generation system is to rank all the edited music-compliant professional-like videos. The cost ranking module computes the average cost (AC) for each edited UGV (i.e., each music-compliant professional-like video) as

$$AC = \frac{1}{M} \sum_{t \in \text{ time indice of selected UGV segments}} OC_t, \quad (9)$$

where M is the length of the candidate music track. For the video segment associated with the local path at 45 degrees, the OC_t equals to the TC_t, while for the segment associated with the local path at zero degrees, the OC_t is given by (8), as shown in Figure 2. Note that the lengths of the generated MVs for a UGV are different, depending on the lengths of the accompanying music

tracks. The generated MVs are ranked in ascending order of the average costs, and the top one is regarded as the best MV recommended to the user. Since every CC_t contributes an additional cost to the average cost, our ranking strategy will favor the music track that makes the edited UGV contain less concatenation points.

4 EXPERIMENTS

To evaluate the effectiveness of the proposed MV generation framework, two scenarios are considered in the experiments. The first scenario concerns soundtrack recommendation only. Given a short video (e.g., about the same length of a general popular song or shorter), the goal is to find a ranked list of music candidates for the video. Therefore, there is no need to perform DTW video-music alignment and video editing in Figure 1, and cost ranking is replaced by a simple similarity ranking scheme. Specifically, the video is paired with each music track from the target music database to form a testing pair. After video segmentation, the MDNN is applied to obtain the pseudo song for each video segment. The PDSM metric is then applied to measure the similarity between the acoustic features of the pseudo song and the music segment, in a segment-by-segment manner. For each testing pair, the total similarity score is summed from the similarity scores of all video segments. Finally, the music tracks are ranked in descending order of the similarity scores, and the top one is used to generate the MV. By contrast, the second scenario executes the complete MV generation process proposed in this paper. That is, the first scenario is a simplified case of the second scenario.

For the first scenario, we performed experiments on a set of OMVs downloaded from YouTube. 265[1] 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 training 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]. In addition, each training OMV was also annotated with a slow-rhythm label or a fast-rhythm label by the authors. The remaining 200 OMVs were used for testing.

For the second scenario, five long UGVs (covering the topics of proposal, wedding, graduation, and travel) collected from YouTube were used as the test videos. The MDNN, the PDSM metric, and the PDCC metric were trained by the same annotated 65 OMVs. Specifically, 8,176 segments (2 seconds long) were generated for MDNN learning; 21,967 pseudo song-music segment pairs were generated for training the PDSM metric; while 14,734 preceding-succeeding segment pairs were generated for training the PDCC metric.

For music analysis, 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. Uniform segmentation was applied to each music track. We then extracted the mean features from the audio frames corresponding to a music segment as the 46-dimensional segment-based acoustic features. For video analysis, the frame-based color themes and motion intensities were extracted as the 8-dimensional low-level visual features [21,22]. The minimum, mean, and maximum values from the frame-based low-level visual features in each video segment were extracted to generate the 24-dimensional segment-based low-level visual features. For high-level visual feature extraction, inspired by the recent success in learning convolutional neural networks (CNNs) for object classification [23–25], we used the 16-layer VGG-Net [25] to obtain the high-level visual features. The VGG-Net was trained on the ImageNet large-scale visual recognition challenge 2012 (ILSVRC-2012) dataset [24]. This collection includes 1.3 million images over 1,000 object categories. Along with the forward propagation in VGG-Net, we extracted 1,000 features from the output layer for each input image. The mean values from the frame-based high-level visual features in each video segment were extracted to generate the 1,000-dimensional segment-based high-level visual features. Finally, the 1024-dimensional segment-based visual features were used as the representation of a video segment. The MDNN contained 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 the PDCC metric, we used a DNN with 3 hidden layers, each with 230, 120, and 60 neurons, respectively. The size of mini-batch for the stochastic gradient descent algorithm was set to 1. For all the three 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.

4.1 The First Scenario: Soundtrack Recommendation

For the evaluation of soundtrack recommendation, we compared the proposed pseudo-song-based matching framework with the state-of-the-art DEMV-matchmaker$_{com}$ framework in [8]. We applied the setting of DEMV-matchmaker$_{com}$ framework in [8] to the proposed framework, i.e., only the acoustic features, low-level visual features and emotion labels were used in both systems. In the experiments, the video of each test OMV was used in turn to search for the best matched music from the music tracks of the 200 test OMVs, and the one corresponding to the test video was regarded as the ground truth. The ranking accuracy [4] defined as

$$Ranking\ Accuracy = 1 - \frac{rank(g) - 1}{|C| + 1}, \tag{10}$$

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.

[1] The video links are available at https://sites.google.com/site/automvgeneration/our-dataset

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

DEMV-matchmaker$_{com}$ [8]	Pseudo-song-based Matching
0.6519	0.7542

Table 2: Average ranking accuracy of the pseudo-song-based framework with different MDNN input/output (I/O) settings

MDNN$_{set1}$	MDNN$_{set2}$	MDNN$_{set3}$	MDNN$_{set4}$
I: LVFs **O**: AFs (main task) and E (second task)	**I**: LVFs+HVFs **O**: AFs (main task) and E (second task)	**I**: LVFs **O**: AFs (main task) E+MS (second task)	**I**: LVFs+HVFs **O**: AFs (main task) E+MS (second task)
0.7542	0.7825	0.7796	0.8061

LVFs: low-level visual features, HVFs: high-level visual features, AFs: acoustic features, E: the emotion label, MS: the music style label

The results in Table 1 demonstrate that the proposed framework outperforms DEMV-matchmaker$_{com}$. One reason is that DEMV-matchmaker$_{com}$ did not consider the relation between music and video modalities in the respective emotion recognition model construction. It might 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. The multi-task deep neural network (i.e., MDNN) used in the proposed framework seems to indeed model the relation among the acoustic (music) features, visual (video) features, and emotion labels. The results may also be attributed to the relatively higher risk in the DEMV-matchmaker$_{com}$ framework, since it needs two mapping processes (i.e., mapping the music and video modalities into the same emotional space separately), while the proposed framework conducts only one mapping process (i.e., mapping the video into the music space). Even DEMV-matchmaker$_{com}$ has used a similarity learning metric (i.e., the EDSM metric) [8] to alleviate the impact of emotion recognition errors, the performance is still limited. Compared to the EDSM metric, the PDSM metric learned with additional less-positive training examples may have a better generalization ability. Overall, the proposed framework pushed ahead the rank of ground truth music by approximately 20, compared to DEMV-matchmaker$_{com}$. The average ranking accuracy was improved from 0.6519 to 0.7542.

The next set of experiments was conducted to evaluate the MDNN. We investigated whether the high-level visual features (i.e., the object representations) and the label of slow-rhythm/fast-rhythm music style could further boost the average ranking accuracy for the proposed framework. The results are shown in Table 2. Compared to MDNN$_{set1}$, which used only the low-level visual features as input, MDNN$_{set2}$ improved the average ranking accuracy by additionally considering the high-level visual features in the input layer. The improvement is expectable because the shot/scene can be represented as a set of object composition [26,27]. In fact, the literature has shown that, in video production (such as a film or OMV), an experienced director is good at applying different shots or scenes to match music to convey the emotion, ideas, and art [28,29]. Accordingly, the segment-based object representations improve the prediction of segment-based pseudo acoustic features, thereby improving the average ranking accuracy. MDNN$_{set3}$ outperformed MDNN$_{set1}$ by additionally including the music style label (i.e., the slow-rhythm/fast-rhythm label) in the output layer. Since the music style and the acoustic features are highly correlated, through the MDNN joint learning architecture, they can transfer knowledge to each other in the common internal (hidden layer) representation to help the MDNN generalize better in pseudo acoustic feature prediction for the unseen video. Finally, by jointly considering all the visual, acoustic, and semantic information, MDNN$_{set4}$ achieved the best average ranking accuracy. It pushed ahead the rank of ground truth music by approximately 30, compared to DEMV-matchmaker$_{com}$. The average ranking accuracy was improved from 0.6519 to 0.8061.

Subjective evaluation[2] in terms of 5-point mean opinion score (MOS) was conducted on 5 MV sets. Each MV set contained the original official MV (the ground truth) and the MVs generated by DEMV-matchmaker$_{com}$ and the proposed framework with MDNN$_{set4}$. The three MVs were provided in a random order. After viewing each MV, the subject was asked to rate a MOS for the indicator "whether the music track matches the video?". Each MV was evaluated by 17 subjects (recruited from the authors' laboratory and university). The average MOS over all MVs and subjects is shown in Figure 3. It is clear that the proposed framework outperforms DEMV-matchmaker$_{com}$. The results reveal that modeling the relation among music, video, and semantic annotations (i.e., emotion and music style) and considering the high-level visual features can indeed generate more attractive MVs that enhance subjects' viewing and listening experiences. The results also show that the MOS of the MVs generated by the proposed framework is quite close to that of the ground truth MVs, which is really encouraging.

4.2 The Second Scenario: Simultaneous Soundtrack Recommendation and Video Editing

For the evaluation of simultaneous soundtrack recommendation and video editing, we compared the complete pseudo-song-based MV generation framework (cf. Figure 1) to its simplified version. For both systems, MDNN$_{set4}$ was used for pseudo acoustic feature prediction, and DTW with a PDSM metric was applied to align the UGV with the candidate music track by evaluating the similarity between the pseudo acoustic features of the segments of the UGV and the acoustic features of the segments of the music track. The PDCC metric for video editing was not applied in the simplified version. Specifically, for the simplified version, after obtaining the time-aligned path for a UGV and a candidate music track, the video editing stage kept the segments of the UGV that correspond to the local paths at 45 degrees and selected the segment with the minimum target cost (i.e., the reciprocal of the score of the PDSM metric) from each local path at zero degrees to generate the edited MV. The sum and average of the target costs

[2] The MOS results and 5 MV sets for the first scenario are available at https://sites.google.com/site/automvgeneration/home/senario1-demo

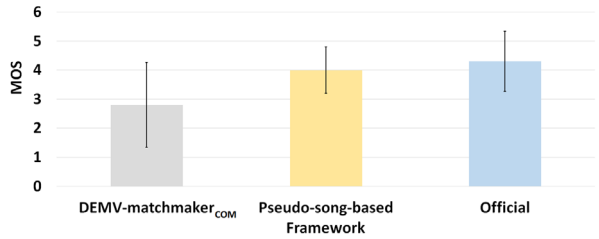

Figure 3: Results of the subjective MOS test for the first scenario (i.e., soundtrack recommendation).

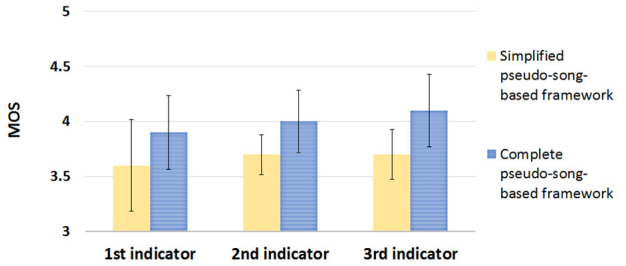

Figure 4: Results of the subjective MOS test for the second scenario (i.e., simultaneous soundtrack recommendation and video editing).

over all segments (including the local paths at zero and 45 degrees) for each edited video were then calculated. Finally, the generated MVs were ranked in ascending order of the average target costs, and the top one was regarded as the best MV recommended to the user.

For the second scenario, since there is no ground truth music track for a UGV, objective evaluation is not possible. Therefore, we performed a subjective MOS test on 5 long UGV sets[3]. A subject was asked to watch the entire original UGV in order to understand the whole story of the video. Then, the subject was asked to rate two MVs generated automatically for the UGV: one was generated by our complete system as shown in Figure 1, and the other was generated by the simplified system. The two MVs were shown to the subject in a random order. The subject was asked to rate the MOS for three indicators: (1) Does the recommended music track match the UGV? (2) Does the edited video show the progress of the story in a professional way? (3) Does the generated MV show a music-compliant professional-like video? Each edited UGV was evaluated by 15 subjects. The average MOS scores for the three indicators over all MVs and subjects are shown in Figure 4. It is clear that the complete system outperforms the simplified system in all indicators. The result with respect to the first indicator clearly demonstrates that, by further considering the PDCC metric (i.e., the concatenation cost) in video editing, the complete system can indeed recommend a more matched music track for a UGV than the simplified system. In other words, the concatenation cost given by the PDCC metric is crucial to not only video editing but also soundtrack recommendation, because the concatenation cost will affect the ranking of the music tracks. In terms of the second indicator, the result indicates that, by applying the PDCC metric in the complete system to select the segments of the UGV, the resulting edited video can show a better progress of the story, compared to the simplified system without using the PDCC metric. Such a result confirms that the PDCC metric can learn, to some extent, the visual storytelling (i.e., the concatenation of segments/shots/scenes) of OMVs. Finally, the strong average MOS score confirms that our complete system performs well and can generate appealing music-compliant professional-like videos with better viewing and listening experiences.

[3] The MOS results, 5 UGVs and generated MVs for the second scenario are available at https://sites.google.com/site/automvgeneration/senario2-demo

5 CONCLUSIONS AND FUTURE WORK

In this paper, a novel content-based MV generation system is proposed based on semantic-oriented pseudo song prediction, matching, and video editing. The proposed system addresses the soundtrack recommendation and video editing challenges simultaneously. Two scenarios have been considered in the experiments, including soundtrack recommendation and simultaneous soundtrack recommendation and video editing. For the first scenario, the proposed pseudo-song-based matching framework outperforms the state-of-the-art DEMV-matchmaker$_{com}$ framework in both subjective and objective evaluations. The results have demonstrated the positive effect of exploring the correlation among music, video, and semantic annotations to automatic music video generation and the ability of the PDSM metric to alleviate the impact of the prediction errors in the pseudo acoustic features on the similarity measure. For the second scenario, the results of subjective evaluation have also demonstrated that, by further considering the PDCC metric, our system can not only recommend a matched music track to a UGV but also edit the UGV into a music-compliant professional-like video according to the matched music track. The generated music video is generally satisfactory and can enhance human viewing and listening experiences.

To further improve the quality of the generated music video, we will investigate how to predict lyrics from the video content in our future work. While a lot of video description tasks have been actively explored, such a task is nontrivial because the lyrics of a music track usually contain a lot of metaphor rather than just declarative sentences. However, we believe that if the pseudo lyrics (cf. the pseudo song in this study) can be correctly predicted from the video content, the selected music track will be more suitable or accurate by additionally considering the matching between the pseudo lyrics of the video and the lyrics of the music track. We will also explore the correlation among the music, video, lyrics, and semantic annotations. In addition, benefiting 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 as well.

ACKNOWLEDGMENTS

This work was partially supported by the Ministry of Science and Technology of Taiwan under Grant: MOST 105-2221-E-001-012-MY3.

REFERENCES

[1] X.-S. Hua, L. Lu, and H.-J. Zhang. 2004. Automatic Music Video Generation Based on Temporal Pattern Analysis. In *Proc. ACM Multimedia (MM)*.

[2] J. Wang, E. Chng, and C. Xu. 2006. Fully and Semi-Automatic Music Sports Video Composition. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*.

[3] J. Wang, E. Chng, C. Xu, H. Lu, and Q. Tian. 2007. Generation of Personalized Music Sports Video Using Multimodal Cues. *IEEE Transactions on Multimedia* 9, 3, 576–588.

[4] F.-F. Kuo, M.-K. Shan, and S.-Y. Lee. 2013. Background Music Recommendation for Video Based on Multimodal Latent Semantic Analysis. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*.

[5] J.-C. Wang, Y.-H. Yang, I.-H. Jhuo, Y.-Y. Lin, and H.-M. Wang. 2012. The Acousticvisual Emotion Gaussians Model for Automatic Generation of Music Video. In *Proc. ACM Multimedia (MM)*.

[6] R. R. Shah, Y. Yu, and R. Zimmermann. 2014. ADVISOR–Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *Proc. ACM Multimedia (MM)*.

[7] J.-C. Lin, W.-L. Wei, and H.-M. Wang. 2015. EMV-Matchmaker: Emotional Temporal Course Modeling and Matching for Automatic Music Video Generation. In *Proc. ACM Multimedia (MM)*.

[8] J.-C. Lin, W.-L. Wei, and H.-M. Wang. 2016. DEMV-Matchmaker: Emotional Temporal Course Representation and Deep Similarity Matching for Automatic Music Video Generation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[9] K. Friston. 2010. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience* 11, 2, 127–138.

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

[11] Q. Summerfield. 1987. 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., 3–51.

[12] D. Chen and B. K.-W. Mak. 2015. Multitask Learning of Deep Neural Network for Low-Resource Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 7, 1172–1183.

[13] R. Xia and Y. Liu. 2015. A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space. *IEEE Transactions on Affective Computing* 8, 1, 3–14.

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

[15] M. Legát and J. Matoušek. 2010. Collection and Analysis of Data for Evaluation of Concatenation Cost Functions. In *Proc. International Conference on Text, Speech and Dialogue (TSD)*.

[16] N. P. Narendra and K. Sreenivasa Rao. 2012. Syllable Specific Unit Selection Cost Functions for Text-to-Speech Synthesis. *ACM Transactions on Speech and Language Processing* 9, 3, 5:1–24.

[17] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1, 18–31.

[18] M. Soleymani, Joep J. M. Kierkels, G. Chanel, and T. Pun. 2009. A Bayesian Framework for Video Affective Representation. In *Proc. Affective Computing and Intelligent Interaction and Workshops*.

[19] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng. 2012. The Acoustic Emotion Gaussians Model for Emotion-Based Music Annotation and Retrieval. In *Proc. ACM Multimedia (MM)*.

[20] O. Lartillot and P. Toiviainen. 2007. A Matlab Toolbox for Musical Feature Extraction From Audio. In *Proc. International Conference on Digital Audio Effects* (*DAFx*).

[21] X. Wang, J. Jia, and L. Cai. 2013. Affective Image Adjustment with A Single Word. *The Visual Computer* 29, 11, 1121–1133.

[22] H.-W. Chen, J.-H. Kuo, W.-T. Chu, and J.-L. Wu. 2004. Action Movies Segmentation and Summarization Based on Tempo Analysis. In *Proc. International Workshop on Multimedia Information Retrieval (MIR)*.

[23] A. Krizhevsky, I. Sutskever, and Geoffrey E. Hinton. 2012. Imagenet Classification with Deep Convolutional Neural Networks. In *Proc. Neural Information Processing Systems (NIPS)*.

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li. 2015. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3, 211–252.

[25] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*.

[26] J. Chen, Y. Han, X. Cao, and Q. Tian. 2013. Object Coding on The Semantic Graph for Scene Classification. In *Proc. ACM Multimedia (MM)*.

[27] W.-L. Wei, J.-C. Lin, T.-L. Liu, Y.-H. Yang, H.-M. Wang, H.-R. Tyan, and Mark H.-Y. Liao. 2017. Deep-Net Fusion to Classify Shots in Concert Videos. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[28] D. Andrews. 2011. *Digital overdrive: communication & multimedia technology*. Digital Overdrive.

[29] G. Mercado. 2010. *The filmmaker's eye: learning (and breaking) the rules of cinematic composition*. Taylor & Francis.