Multimedia Tools and Applications
https://doi.org/10.1007/s11042-019-08435-x

A music-driven system for generating apparel display video

Hui Zhang[1] · Kejun Zhang[1] · Yingping Cao[1] · Jun Zheng[1] · Xiaoyi Huang[1] · Changyuan Yang[2] · Lingyun Sun[3]

Received: 13 March 2019 / Revised: 7 August 2019 / Accepted: 1 November 2019 /
Published online: 07 December 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract
Video plays a great important role in online apparel sales, which is a vital tool for publicity and to provide consumers with space of imagination. However, as the apparel market rapidly updates in large amounts every day, creating videos for fast increasing clothes can be challenging and labor-consuming. Considering this, we present ApVideor, a music-driven video generation system customized for displaying clothes. This system consists of two main modules: music recommendation module and audio-visual synthesis module. The former assists users in searching background music that matches the apparel style, while the latter combines the audio and visuals into a video by music-driven approaches. Our user study suggests that this system makes the video creation process significantly easier and faster than manual creation. Meanwhile, the viewer test suggests that apparel-displaying videos created using our system are of comparable quality to those created manually by people who have worked with video editing.

Keywords Music driven approach · Music recommendation · Transition detection · Video generation · Apparel-displaying

1 Introduction

It is widely recognized that video is an excellent vehicle of presenting merchandise, especially for online shopping. When consumers shop online, they are likely to purchase the outerwear with high quality images, yet a more immersive, multimedia shopping experience with video content would make they even respond better. Statistical data from *Multichannel*

✉ Kejun Zhang
zhangkejun@zju.edu.cn

Extended author information available on the last page of the article.

Merchant[1] have shown that 71 percent of consumers consider video as the best option for explaining product features. 65 percent of consumers say they can imagine a product better if they have watched a corresponding video.

When it comes to a real shopping scenario, though the sellers agree on the necessity of the video representation, they may not choose to make display video for their clothing items within budgetary and schedule constraints. Creating videos for rapidly changing clothes is much time-consuming and labor-consuming. It is not surprising to see that a large percentage of the clothing items in *TaoBao.com* do not have video showcase.

Currently, there are a few large e-commerce sites which have launched their own commodity video generation tools such as *Lightning* of *JD.com*. However, these tools can only serve the items on their platform, and the internal algorithm model is not clear. On the other hand, the automatic generation of apparel display video has not been studied in academic community. This calls for research support for online apparel sellers in terms of video making.

To create video representations for apparel, people are required to process complicated multimedia materials. A typical workflow to create such videos using video editing software is shown as follows:

1. Shoot some visual materials for the clothing;
2. Look for suitable background music;
3. Find appropriate transition timestamp and choose corresponding rendering effects;
4. Use video tools to arrange the visual and acoustical materials together.

It is interesting to note that though clothing items on current e-commerce sites do not necessarily have video, they do have pictures (usually seen at "commodity details"). Therefore, it is possible to utilize these pictures to create videos automatically.

However, choosing the right music is a very critical issue since music is not just a part of video, it also has an important impact on the shopping experience. Previous studies on the effect of music on shopping experience has shown that music is of great significance to consumers' emotion, purchase intention and shopping behavior. Judy I. Alpert et al. [1] suggested that music has both theoretical and practical implications on advertising as a stimulus to emotional response and the degree of influence as well as the direction of desirable may vary in different situations. For example, people may response more positively to sad emotional music on purchasing intentions than happy or silent ones in the case of buying a greeting card. While in amusement parks, cheerful background music is much more effective in promoting advertising than sad one. Jochen et al. [21] stated that music may affect the mood and behavior of consumers. It has been found that the customers may spend much more time and money in a supermarket playing low-tempo music. While in fashion retail store, high-tempo music may increase customer's shopping duration and intentions. It is worth noting that not every music can trigger positive shopping behavior, a comprehensive consideration should be given to store type, brand type, music style, emotion conveyed by music, user type and other factors. Only when playing the right music for the right item can music effectively arouse consumer's emotions and promote their purchase intention.

On the other hand, music can also serve as a cue to change the visual scene. Arranging pictures according to the fluctuation of music is conductive to the perceptual synchronization between human hearing and visual sense. There are a few guidelines such as

[1] http://multichannelmerchant.com/mcm-outlook-2013-research-reports/

cut-to-the-beat in the video production industry which can be automated by computers using music pattern recognition approaches.

Taking all these factors into consideration, we intend to propose a music-driven method for developing video auto-generation system which could automate step 2, 3 and 4 in video creation process. In this article, we have implemented a music-driven video maker, ApVideor [2], to support apparel sellers in making video clips. This system assists users by recommending music matching the apparel genre (alleviating workload in step 2 above), and then automatically using uploaded clothing visuals to synthesize video in harmony with background music (alternative to work in step 3 and step 4).

The major contributions of this paper are several folds: We propose a novel and effective procedure to automatically generate apparel display videos. As part of this procedure, we describe a semantic-analysis based approach for providing a mapping between the apparel style and music attribute space. This allows us to effectively recommend appropriate background music for the automatically-generating video. Meanwhile, in this part, we have collected a new dataset from e-commercial website, which contains 14257 soundtracks of women apparel and seller-annotated style information, to support the research in this area. In order to achieve the perceptual synchronization between human hearing and visual senses, pictures in video are required to be organized according to certain rules to adapt to the fluctuation of music. To this end, we also introduce a strategy to handle the combination of acoustic and visual materials with audio pattern recognition to allow a better acoustic-visual harmony. We design and conduct a set of user-based experiments to demostrate the usefulness of the music-driven video generation procedure customized for apparel display.

In the following Section 2, this article will show some existing research related to our work. The system interface and essential operations will be introduced in Section 3. Then the basic framework, underlying algorithms and evaluation will be described in the next Section 4. In Section 5, a set of user studies will validate the design of our system and reveal the reasonable quality of the automated video. Finally, in Section 6 we will draw a conclusion and present some possible extensions.

2 Related work

Our work intends to assist apparel video manufacture, mainly by recommending suitable music and implementing audio-visual synthesis automatically. These technologies have attracted numerous efforts from diverse research fields. In this section, we will review related work in each of these areas since our work is inspired by the literature and draws lessons from prior methods.

2.1 Music recommendation

Music recommendation technology has become a hot topic in the research community of Music Information Retrieval (MIR) and is also increasingly highlighted by the music industry (especially in e-commerce, such as *Last.fm*, *Amazon*, *Netflix*, etc.).

The existing mainstream music recommendation methods can be generally divided into two categories, namely collaborative filtering recommendation algorithm [9, 26] and content-based recommendation algorithm [24]. These techniques are commonly applied to

[2] ApVideor: A Video Automaker for Apparels

the recommendation module of music streaming platforms, e.g., *Netease Music* and *Tencent Music*, and require user profile data including the personal information and preferred behaviors like preferences, favorites, downloads, etc. In these cases, the focus of music recommendation techniques is to explore the music preferences of listeners. For example, systems in [2, 6] guess listeners' interests by their operating records, and then recommend the preferential music to them.

However, in the research scene that recommending music for apparel, we are faced with the new task which requires us to build a map from music feature space to target apparel. The available input information of apparel music selection is totally different from user-profile based recommendation algorithms. As far as we know, there is little existing work in customizing music recommendation algorithm for apparel. Yet there are still some research which share correlated perspective with our work.

Chao et al. [5] developed a semantic-driven music recommendation service for digital photo albums to capture relatedness between images and music using text tags as intermediary, which is the essential part of this technology. This is a salutary lesson leading us to explore the relationship between apparel style and music genre, which are both text label of the two targets.

Meanwhile, some researchers were motivated by recent development in affective computing. They have tried mapping the different types of multimedia features into an emotional space [16, 23]. These studies have laid the foundation of X Wu's project [25] which investigated the matching degree between image and music using emotion, and JC Lin's paper [15] which mined the relationship between music and video. Our work also adopts emotional information as bridge.

2.2 Audio-visual synthesis

Music and images have been inevitably combined together in the history of filmmaking since the end of silent-film era (the late 1920s) [3]. The key role of music in producing video has also been emphasized by the music video industry (MTV). One procedure of our system is designed to sequence a stack of visual materials to coordinate with the music (i.e. audio-visual synthesis). A research problem related to audio-visual synthesis is automated video editing (AVE). Numerous contributions have been reported to this topic, which are very inspirational and helpful to our work on video creation.

A critical question in audio-visual integrity is to find appropriate video cuts. The answer to "where to cut" is pretty subjective and differs between people. Nevertheless, this practice still has a few guidelines such as *cut-to-the-beat* [17]. Our work takes some empirical cutting rules to determine transition points as the music changes.

In addition, Z. Liao et al. [14] proposed a music-driven imagery way to produce a video montage. Given a piece of music and a set of video clips, their system is able to analyze the video and music content and output a composition. The composition principles addressed in their formulation provide us a direction to estimating transitions of the result video. However, Z. Liao's system requires the background music to be in *MIDI* format, which imposes restrictions on common music resources. Hua's research [13] in AVE has also studied the association between acoustical information and visuals. Aiming at more professional-looking results, he developed an approach for extracting a temporal structure, beats and tempos from music so that the video segments were able to match the content of the incidental music.

Another question is related to transition effect, which is an important tool connecting visuals naturally. Nam's paper [18] presented a scheme that focused on analyzing the characteristics of the underlying special edit effects in video sequences. Wang et al. [22] used multi-modal cues to generate personalized music sports video. The automatic system implemented in [12] defined a set of cutting rules to emphasize the repetitive patterns of the background music. All of this research adopts corresponding methods to choose proper transition effects for connecting visual sub-clips. In this article, we summarize some rules (see details in Section 4) empirically for choosing transition effects like aforementioned reference.

3 ApVideor interface

ApVideor is a web application based on *Django Development Toolkit*[3] with our algorithms running as the backend in *Python*. In this section, we will present the interfaces and describe essential operations of ApVideor.

The interaction of ApVideor system is designed based on simplicity and usability. It serves people who do not have much editing abilities, especially in e-commerce, and helps them to edit video efficiently. Users are simply required to select the style of clothing they want to display and upload the visuals (pictures and/or video segments), and then ApVideor will recommend background music and generate video for them. The interface of tool mainly consists of two parts, namely navigation bar and operation module (see in Fig. 1).

* Navigation Bar. There is a "return" channel to homepage on the left of navigation bar and a user manual is on the right. The design of navigation bar satisfies general internet use habits, which can reduce learning cost of users effectively.
* Operation Module. Fitts' Law indicates that the time required for moving from a position to another is related to the distance between them and the size of each. The longer the distance is, the longer time it takes, while the larger the size of target is, the shorter time it takes. To ensure the tool's efficiency and usability, the operation buttons of appropriate size are clustered in a relatively fixed and centralized area.

Figure 2 shows the snapshots of our demonstration and represents the operational flow of the four key pages of the ApVideor system.

(a) Select the style. In order to reduce user's operational costs, ApVideor provides the drop-down menu for users. In this page, users need to select the corresponding style of the commodity, which is the input item for music recommendation.

(b) Upload visual materials. Card design has been a common interaction mode of Internet products recently, which can maintain the visual unity of elements. As a pseudo-element, card conveys a clear operability like being stacked, covered, moved and crossed out. The interactive behavior of image and video materials uploaded by users is presented in this way.

(c) Recommended music. Users are able to listen to the recommended music with the motion design of music insight to get high-quality audio-visual experience. Music pieces are displayed in the order of high-to-low matching degree. Only one song can be played at a time, and users can switch songs by pressing button on the left or right according to their subjective feelings.

[3] https://www.djangoproject.com/



Fig. 1 Navigation bar and operation module

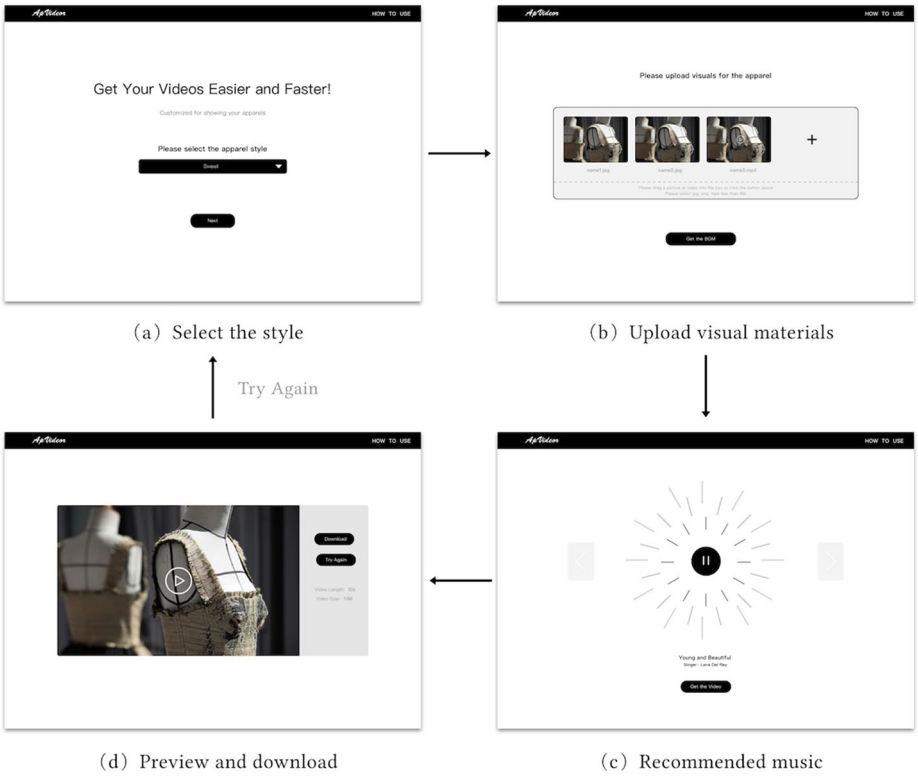


Fig. 2 Key pages of ApVideor user interface

(d) Preview and download. ApVideor renders the video clip quickly after finishing aforementioned operations. For the consideration of actual user needs, it supports operations like video preview, download, regenerate et al. The buttons are placed on the right side of the panel in order to satisfy the usage habit of most people.

4 Algorithms and evaluation

The basic framework and underlying algorithms will be introduced in this section. All the project codes and datasets used in this article are available online[4].

The basic architecture of ApVideor system is shown in Fig. 3. The apparel information, including style and visual materials, is inputted into music recommendation module and audio-visual synthesis module respectively. The former module searches the music collection for suitable background songs and the latter one makes a composition of the recommended music and the visual materials. The final output is a display video for the apparel.

4.1 Music recommendation module in ApVideor

Prior research in psychology and marketing has shown that music as frequently used practice of mood induction has significant impact on purchase intentions [1]. Especially, while the ability of happy or sad musical content to influence mood seems likely, the degree of influence and the desirable direction may vary in different advertised items. Therefore, in order to find suitable music for specific apparel, we should analyze the relationship between music emotion and apparel attributes. In addtion, music genre and apparel style naturally share similar semantics. For example, video editors usually match punk-style clothes with punk music. So we consider genre as another dimension of music attribute space to build a mapping between music and apparel. Furthermore, prior research in consumer behavior has shown that specific background music selections along dimension of increasing popularity could positively affect responses to advertised products [19]. As a result, social data collected from online music platforms such as favorites and followers can be used as a ranking metric for apparel music recommendation. Based on the above research conclusions, we constructed the architecture of our music recommendation module in Fig. 4.

To build the apparel music recommendation module of ApVideor, we first captured the clothing data from one of the most representative e-commercial websites, namely *TaoBao.com*. The captured data include apparel videos and corresponding style information. Videos without background music were filtered out and soundtracks of the remaining ones were extracted to perform analysis. In total, 14257 style-audio pairs were collected to form the dataset. To support the research in this area, we have also published this brand new dataset online which contains background soundtracks of women apparel and seller-annotated style information. It can be downloaded free from our website.

Background music was extracted from the videos and performed analysis. In this procedure, two main tasks are Music Emotion Recognition (MER) and Music Genre Recognition (MGR). Specifically, music emotion here is represented by *Valence* describing the positiveness conveyed by the music [27]. Music with high valence sounds more positive (e.g., happy, cheerful, euphoric), while that with low valence sounds more negative (e.g., sad,

[4] https://deryy.github.io/ApvideorDemo/page.html

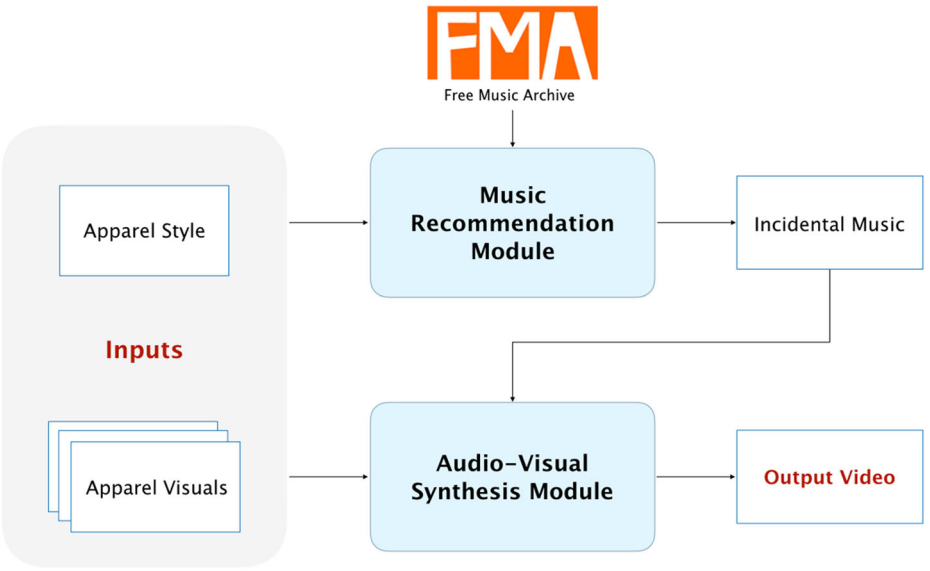


Fig. 3 Structure of ApVideor system

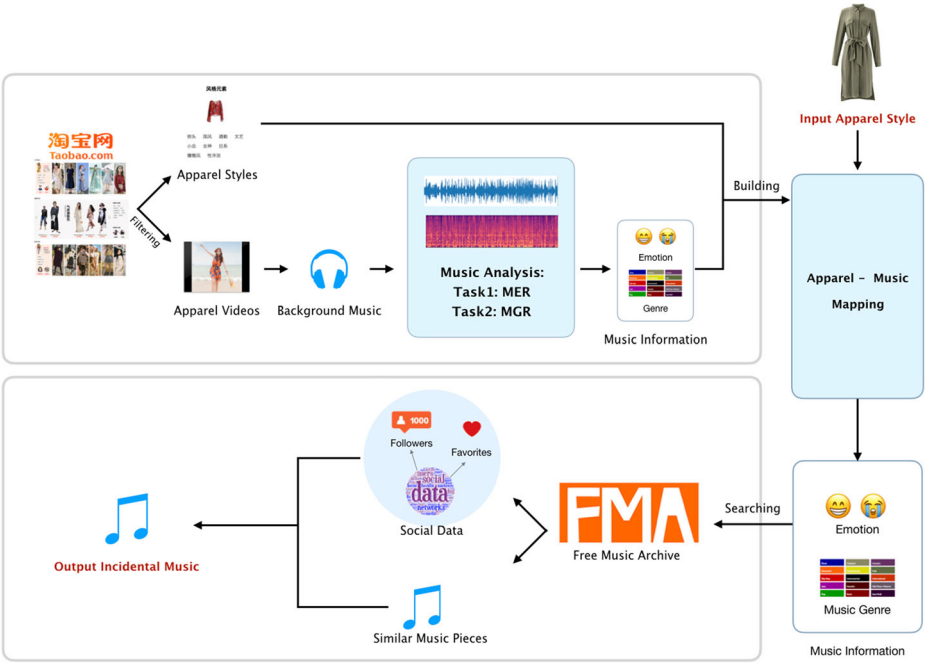


Fig. 4 Music Recommendation Module in ApVideor

depressed, angry). MER task in this procedure aims to classify the music into two categories (positive/negative). With regard to MGR task, we adopt 8 categories of music genre (*hip-hop, pop, rock, folk, experimental, electronic, international, instrumental*), which are used to constitute the balanced FMA dataset [8]. Two classifiers were trained based on data from FMA dataset to achieve these two tasks.

Yandre MG's work [7] has shown outstanding performance in music classification. Inspired by his work, our study adopts Convolutional Neural Network (CNN) to find the association between audio signal and its high-level semantic concepts. In particular, our study has transferred audio signals into spectrograms (images of 256 pixel height, $50 \times d$ width, d is the duration of music) using discrete Fourier transform. FMA dataset contains 8000 music tracks of 30 seconds. We transferred each track into spectrogram image of 256×1500 pixels. It is worth noting that the collected videos have different lengths, so the duration of background music is not fixed. To deal with this problem, we split the image into patches of 256×150 pixels (sub-images of 3-second music segments with no overlap).

Afterwards, as shown in Fig. 5, we trained two VGG-16 [20] networks with batch normalization as classifiers to recognize music genre and emotion respectively. A VGG network of 16 weighted layers is constructed by 13 convolutional layers and 3 fully-connected layers. We implemented the network based on *Pytorch Torchvision Library*[5] and chose stochastic gradient descent (SGD) as optimizer. SGD updates a parameter θ for each training example $x^{(i)}$ and label $y^{(i)}$:

$$\theta = \theta - \eta \cdot \nabla_\theta J\left(\theta; x^{(i)}; y^{(i)}\right) \tag{1}$$

where η is the learning rate. Cross-entropy loss was adopted to measure the deviation between the predicted result and the target value:

$$\text{Cross-entropy Loss} = -\sum_{c=1}^{M} y_c \log\left(p_c\right) \tag{2}$$

M is the number of classes. y_c denotes true label as one-hot encoding and p_c is the predicted probability of the final soft-max layer for class c. The resulted models are able to predict genre and emotion labels from music with 69 percent and 74 percent accuracy respectively.

The recognized (emotion, genre) labels of music enable us to build a bridge between apparel style and music using semantic tags. A statistical analysis was performed on the 14257 pairs of style-audio data. For each apparel style (31 categories of apparel style in total, e.g., sweet, unisex, punk), the two sets of (emotion, genre) with the highest frequency will be used to search the music collection for corresponding songs which have the same music attributes. The searched results are ranked in descending order of hotness (estimated by the number of favorites). Finally, ApVideor displays the top five songs as the recommended ones.

To evaluate the performance of our recognition model, we compared the prediction results with those of multiple standard classifers trained with mainstream audio features which are commonly used in music information retrieval (MIR) research. Specifically, the 518-dimensional feature set consists of chroma, tonnetz, MFCC, spectral centroid, spectral bandwidth, spectral contrast, spectral rolloff, RMS energy and zero-crossing rate (see

[5] https://pytorch.org/docs/stable/torchvision/models.html

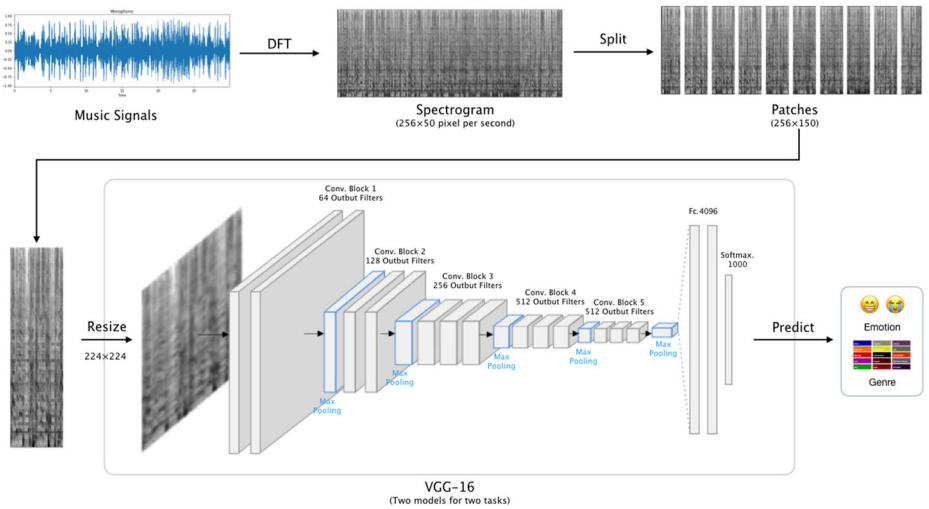


Fig. 5 Implementation of MER and MGR in ApVideor System

Table 1). Each of them (except zero-crossing rate) is calculated on windows of 2048 samples spaced by hops of 512 samples. The dataset was splitted into 10 folds. 9 folds composed the training set and 1 fold was used as the test set.

Table 2 reports accuracies for muisc genre recognition and music emotion classification with nine standard classifiers and our model. To be specific, we employed k-nearest neighbors (kNN) with k = 200, support vector machines (SVM) with multiple kernel functions including polynomial, linear and radial basis function, decision tree (DT) with $depth = 5$, random forest (RF) with $depth = 5$, multilayer perceptron (MLP) with 100 and 200×50 hidden neurons, Gaussian naive bayes (GaussianNB), adaptive boosting (AdaBoost) as well as quadratic discriminant analysis (QDA). It can be seen that our method which constructs convolutional neural network based on audio spectrogram achieved best performance both in MGR and MER tasks.

In general, the existing mainstream music recommendation methods can be divided into two categories, namely collaborative filtering recommendation algorithm [9, 26] and content-based recommendation algorithm [24]. These techniques are commonly applied to the recommendation module of music streaming platforms, e.g., *Netease Music* and *Tencent Music*, and require user profile data including the personal information and preferred behaviors like preferences, favorites, downloads, etc. However, in the research scene that recommending musisc for apparel, we are faced with the new task which requires us to build a map from music feature space to target apparels. The available input information of apparel music selection is totally different from user-profile based recommendation algorithms, which leads to the incomparability of our methodology with other mainstream methods. Therefore, in order to evaluate the performance of our music recommendation module, we invited human users to carry out this task in Section 5 and compared the results through user study.

Table 1 Audio feature set

Feature	Dimension	Content
Chroma	84	12-dimensional chroma vector [10] and their mean, standard deviation, skew, kurtosis, median, minimum and maximum.
Tonnetz	42	6-dimensional tonal centroid vector [11] and their mean, standard deviation, skew, kurtosis, median, minimum and maximum.
MFCC	140	20 MFCC coefficients and their mean, standard deviation, skew, kurtosis, median, minimum and maximum.
Spec. centroid	7	the mean, standard deviation, skew, kurtosis, median, minimum and maximum of spectral centroid.
Spec. bandwidth	7	the mean, standard deviation, skew, kurtosis, median, minimum and maximum of spectral bandwidth.
Spec. contrast	49	spectral contrast values corresponds to 7 octave-based frequency sub-bands as well as their mean, standard deviation, skew, kurtosis, median, minimum and maximum.
Spec. rolloff	7	the mean, standard deviation, skew, kurtosis, median, minimum and maximum of roll-off frequency.
RMS energy	7	the mean, standard deviation, skew, kurtosis, median, minimum and maximum of root-mean-square energy.
Zero-crossing rate	7	the mean, standard deviation, skew, kurtosis, median, minimum and maximum of zero-crossing rate.

4.2 Audio-visual synthesis module in ApVideor

As shown in Fig. 6, there are two main steps when solving audio-visual synthesis. Firstly, the system needs to find where to cut (transition points). Then it should choose suitable effects to combine different frames (rendering effect).

In order to decide appropriate positions to align music and visual clips, ApVideor will detect specific rhythm modes in the musical sequence. Four rhythm modes are defined based on music features to prompt transition between images. The definitions of the four rhythm modes are shown in Table 3, which are estimated by *onset* (the moment when a note shows) and *tempo* (beats per minute). We applied *librosa toolkit*[6] to extract onset series and beats. Onset series is represented as $Onset = \{(t_i, s_i)|0 \leq i < N\}$ where t_i is the timestamp of the i-th onset, s_i is the strength and N is the total number of onsets in the

[6] https://librosa.github.io

Table 2 Prediction accuracies of various classifiers in MGR and MER tasks

Classifier	Accuracy of MGR	Accuracy of MER
kNN	.35	.58
SVCrbf	.13	.60
SVCpoly	.56	.65
SVClin	.55	.68
DT	.38	.59
RF	.37	.69
MLP1	.35	.46
MLP2	.45	.60
GaussianNB	.46	.56
AdaBoost	.40	.66
QDA	.43	.60
CNN	**.69**	**.74**

Bold values are the highest accuracies for the two tasks

music. $M_i = \{s_i, ..., s_{i+9}\}$ means 10 continuous onsets and $\{s_i\}^* = Normalization(\{s_i\})$. The tempo, calculated as the frequency of musical beats, is used to divide the music into three types (fast, moderate and slow) with different priority lists of four modes to decide boundaries.

Visuals are arranged in upload order (i.e., user-specified display order). Analysis of visual features has not been performed to arrange images in the first iteration of ApVideor. That is because we have not yet found a suitable model to automatically arrange the pictures. We have tried to show the overall picture first and then the detailed picture. However, due to the great difference between photos shot by different e-commercial merchants, we have not achieved an effective method to accurately distinguish these picture types. This feature will be improved in our following work.

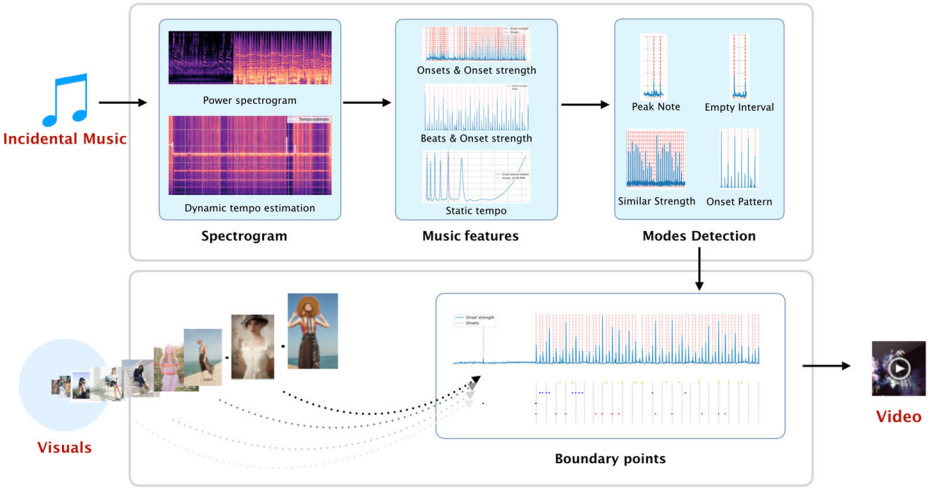


Fig. 6 Audio-visual synthesis module in ApVideor

Table 3 Four rhythm modes and rendering effects

Rhythm mode	Definition
Peak Note	$\{t_i \mid \frac{s_i}{s_{i-1}} > 2\}$
Empty Interval	$\{t_i \mid (t_i - t_{i-1}) \geq 1.8 \times \frac{60}{tempo}\}$
Similar Strength	$\{t_i \mid \|M_i - M_j\| \leq 0.2\}$
Onset Pattern	$\{t_i \mid \{s_i^*, s_{i+1}^*, s_{i+2}^*, s_{i+3}^*\} \in \{\{0,1,0,1\},\{1,0,1,0\}\}\}$

Rendering Effect	Definition
Momentary Cut	$duration < 1s$
Zooming & CF	$1s \leq duration < 3s$
Slow Motion	$3s \leq duration < 4s$
Cross Fade	$4s \leq duration$

After assigning the timestamp to visuals, the system is able to render the video using some effects. According to observations on typical MVs, we have found some most commonly used rendering effects for transition: momentary cut, cross-fade, zooming, shading and slow motion. ApVideor obtains the final music video by connecting the image sub-clips and (or) video sub-clips with aforementioned effects empirically based on transition duration (shown in Table 3).

There are many cases of video generation in academia and we mentioned some similar ones in Section 2. However, these methods were found not able to be compared with our method directly. There are two main reasons (we take Liao's [14] and Hua's research [13] as examples). The first reason lays on different service objects. In Liao's work, a montage video generation method is proposed. Operators are required to provide music and videos, and then rearrange clips cut off from music timelines according to music-driven videos. This work aimed at serving the clip combination of montage technique. In Hua's research, an automated home video editing system is proposed. Similarly, operators need to provide music and several home videos. This system has three stages: content analysis, content selection and composition. The system automatically selects suitable or desirable highlight segments from a set of raw home videos and aligns them with a given piece of incidental music to create an edited video segment to a desired length based on the content of the video and incidental music. The second reason is the different inputs. Both articles generate videos based on video clips. They analyze motion of the video segments and match them with music rhythm, while we insert noncontinuous pictures between music transition points for video generation. In addition, the music analysis of the first article is based on the MIDI file where music tone and intensity are directly stored, while we extract the music information from audio waveform in MP3 file. The visual materials of the second article are home videos with voice inside, so sentence boundary detection is also needed to align the audio track with music rhythm. Considering these reasons, techniques in these papers can not be applied directly to solve our problem.

However, some parts of the existing work can be used for music transition analysis. Two methods of music rhythm analysis are mentioned in Liao's work, namely the typical cut-to-the-beat principle and agglomerative clustering. Cut-to-the-beat means visual scenes should be changed at significant note onsets in the music. Agglomerative clustering is a music analysis method proposed to divide music into multiple segments [14], which is performed bottom-up to form a hierarchical clustering tree. The minimum music segments form the initial set (Liao used the bars in the music as the atomic units for music segmentation). The

pair of consecutive segments with the minimum distance is chosen to merge during each step of clustering and this process will be repeated until the number of notes in all segments has reached a desired segmentation granularity. We reproduced cut-to-the-beat principle and agglomerative clustering method to create videos and compared them with results generated by our audio-visual synthesis module so that we can evaluate the performance of different approaches.

As far as we know, at present, there is no objective evaluation critia to quantify the editing quality of video in the world, so we can not carry out the quantized evaluation of the results. Instead, we recruited humans to evaluate the qualitative performance of our algorithm (Group A) comparing to 2 control groups. Videos in the first group are results of cut-to-the-beat approach (Group B) and the others are results generated using agglomerative clustering (Group C).

This study was performed on eight apparel items and each item had three result videos created by three methods respectively (8×3 result videos in total). 24 participants' ratings were collected. Each of them was asked to watch the above four sets of videos (a set means three results of the same apparel item) and then rate the results in a scale between 1 and 5 where 1 means the worst and 5 means the best.

Finally, each set of videos was rated by 12 participants and each group of results received 96 (8×12) ratings. Table 4 shows average scores by various subpopulations according to information collected from questionnaire. It can be seen that our result holds a notable advantage away from the control groups in both of average rating and standard deviation. We also performed paired t-test on the three groups of results which indicated that our results were rated significantly higher than the results either of cut-to-beat ($t = 2.98$, $p = 1.85 \times 10^{-3}$) and agglomerative clustering ($t = 3.53$, $p = 3.24 \times 10^{-4}$). We also compared the performance of our algorithm with human users in the following section.

5 User study

A set of user studies were designed and conducted to validate our design. The usability test aims to measure whether or not creating videos using our system is easier and faster than manual creation. Then independent viewer tests were performed to compare the quality of videos created using the ApVideor system with the manually created videos.

5.1 Usability test

For validating the design of our system, we recruited 17 video editors from the age of 22 to 24 (9 females and 8 males). The sample size of 17 participants in this test was determined according to the local standards for sample size at CHI (17 for in-person within-subjects

Table 4 Average ratings of results by different methods and subpopulations

Subpopulation	#ratings	Ours	Cut-to-beat	Agglomerative clustering
all	96×3	$\mathbf{3.64 \pm 0.90}$	3.05 ± 1.32	3.29 ± 1.03
female	24×3	$\mathbf{3.58 \pm 0.95}$	3.54 ± 1.35	3.16 ± 1.07
male	72×3	$\mathbf{3.65 \pm 0.88}$	2.89 ± 1.26	3.33 ± 1.01

Bold valuess are the highest average user ratings for each subpopulation

experiment) [4]. We defined video editor as someone who claimed to have work experience in video editing. 8 sets of apparel visuals (8 pieces of clothes randomly selected from *TaoBao.com*) were downloaded as the test materials.

Each participant was asked to manually create two videos for two pieces of clothes respectively (the length of each video was controlled less than 30 seconds). They were allowed to work with any video editing software they liked. After this creation, participants were asked to rate the difficulty level of creating videos on a scale of 0 to 5 (0 means extremely easy and 5 means extremely difficult). The two stages of this task were also rated individually. The first stage refers to finding background music (step 2 in Section 1) and the second refers to synthesizing the music and visuals (step 3 and step 4 in Section 1). After manual creation, participants were required to use the ApVideor system to create videos automatically and rate the difficulty level. The reported ratings and the time taken to finish each task were recorded.

As shown in Fig. 7, the median score of difficulty rated by subjects was 3 for manual video creation (3 for stage 1 and 2 for stage 2) but 1 for creation using ApVideor. The difficulty level of creation using ApVideor was reported to be significantly lower ($t = 3.8996$, $p = 3.626 \times 10^{-4}$). Meanwhile, the time spent on manual creation (2 videos) is also significantly longer than that spent on creation using ApVideor ($t = 13.089$, $p = 1.271 \times 10^{-10}$).

5.2 Viewer test

In order to evaluate the quality of videos created by ApVideor, we recruited 70 subjects to perform a viewer test comparing videos created by ApVideor with those created manually (41 females and 29 males, subjects who have participated in user study were excluded).

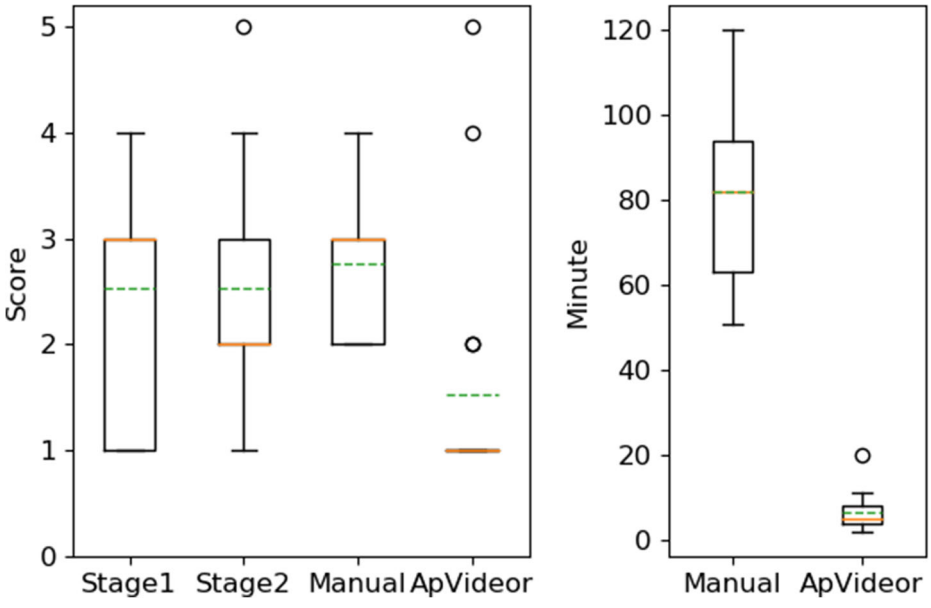


Fig. 7 Level of difficulty for video creation reported by subjects (the left one) and time spent on two tasks (the right one). The orange full lines denote the median value and green dotted lines denote the mean value

Table 5 Average scores of background music (S1) and visual transition (S2), as well as the preference reported by subjects

Type	S1	S2	Preference
ApVideor	**3.65**	**3.52**	**57.71%**
Manual	3.50	3.44	42.29%

Bold values are the better scores for music, transition and preference

We know that image quality is important for video, and making visual material is also time-consuming and difficult. However, the focus of this article is not on capturing or generating the visual assets (we will further study this content in future work) but on helping the sellers who have already prepared necessary introduction pictures to create videos easily and fast. Therefore, the quality of video in this test is defined in terms of corresponding music and visual transition.

Specifically, we required the subjects to rate the level of how well the incidental music matches the apparel style and how well the visual transitions match the music on a scale of 1 to 5 (5 means perfectly match, 4 means moderately match, 3 means acceptable, 2 means not so pleasant, and 1 means entirely unsuitable). We also asked subjects to directly compare the videos created manually with those created by ApVideor.

During the aforementioned process of user study, we have collected a total of 34 pairs of manually created videos as well as corresponding automatically generated ones (17 participants $\times 2$ videos/participant). These videos were randomly divided into 7 groups. We duplicated 1 pair of videos to balance the number of videos in each group (5 pairs). Each set of videos was rated by 10 subjects. This task took each participant about 15 minutes.

As shown in Table 5, the videos created by ApVideor were rated higher than those created manually both in terms of background music and visual transition. Meanwhile, we can see from the same table that, on average, subjects preferred videos created by ApVideor to those created manually.

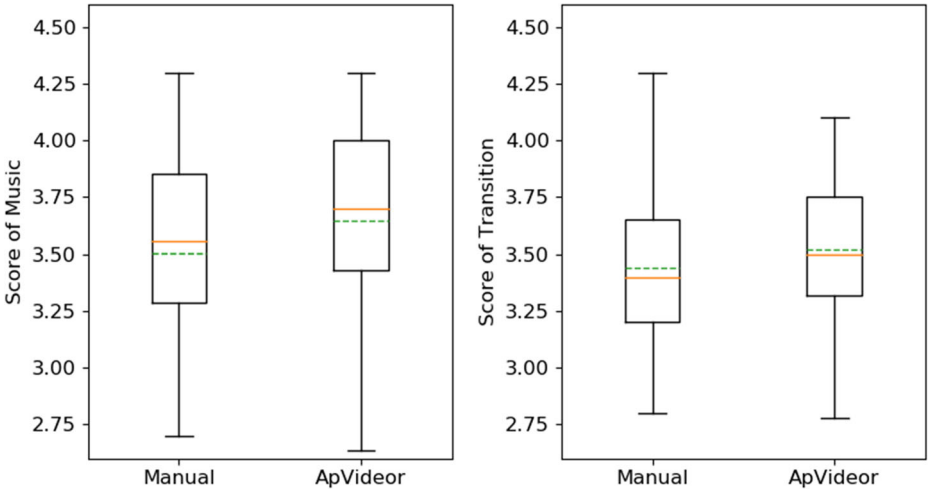


Fig. 8 Scores of videos rated by subjects (in terms of music and transition) on a scale of 1 (entirely unsuitable) to 5 (perfectly match). The orange full lines denote the median value and green dotted lines denote the mean value

Figure 8 shows two distributions of scores rated by subjects for two questions respectively. Since there is no significant difference between these two distributions, we used two-one-sided t-tests (TOST) to perform an equivalence testing. Results have shown that ApVideor can find music that matches the apparel style as well as humans ($t_{(\Delta_L=-0.3)} = 1.95$, $p_{(\Delta_L=-0.3)} = 2.6 \times 10^{-2}$, $t_{(\Delta_U=0.3)} = -5.64$, $p_{(\Delta_L=-0.3)} = 1 \times 10^{-8}$) and produce visual transitions like people who have work experience in video editing ($t_{(\Delta_L=-0.3)} = 2.98$, $p_{(\Delta_L=-0.3)} = 1 \times 10^{-3}$, $t_{(\Delta_U=0.3)} = -5.04$, $p_{(\Delta_L=-0.3)} = 3 \times 10^{-7}$).

In this section, the usability test has validated that creating videos using ApVideor system is easier and faster than manual creation and the results of viewer test have revealed the reasonable quality of the videos created by ApVideor.

6 Conclusion

In this paper, we have presented a music-driven video generation system, ApVideor, which is customized for supporting apparel sellers. The ApVideor system implementes a semantic-analysis based approach for providing a mapping between the apparel style and music attribute space. This allows the system to effectively recommend appropriate background music for the automatically-generating video. Meanwhile, to support the research in this area, a new dataset have been collected from e-commercial website and is now available to download free online. Also, a strategy is proposed to handle the combination of acoustic and visual materials with audio pattern recognition to allow a better acoustic-visual harmony. Our experiment suggests that creating apparel-showing videos using our system can be significantly easier and faster, and the result videos have comparable quality to those created manually. A promising direction for future work is to explore other aspects of multimedia contents that may support sales behavior (e.g., capturing, analyzing and generating visual assets).

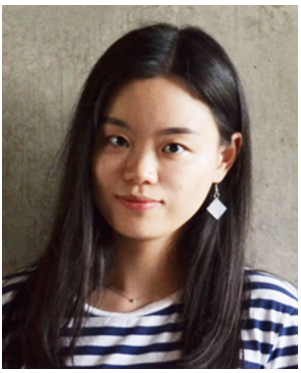
Acknowledgements This project is supported by the Key Project of National Science Foundation of Zhejiang Province (No. LZ19F020002), the National Natural Science Foundation of China (No. 61672451), Provincial Key Research and Development Plan of Zhejiang Province (No. 2019C03137). And the authors would like to thank the Key Laboratory of Design Intelligence and Digital Creativity of Zhejiang Province, Alibaba-Zhejiang University Joint Institute of Frontier Technologies and State Key Lab of CAD&CG.

References

1. Alpert JI, Alpert MI (1990) Music influences on mood and purchase intentions. Psychol Market 7(2):109–133
2. Bu J, Tan S, Chen C, Wang C, Wu H, Zhang L, He X (2010) Music recommendation by unified hypergraph: combining social media information and music content. In: Proceedings of the 18th ACM international conference on Multimedia. ACM, pp 391–400
3. Buhler J, Neumeyer D, Deemer R (2010) Hearing the movies: music and sound in film history. Oxford University Press, New York
4. Caine K (2016) Local standards for sample size at chi. In: Proceedings of the 2016 CHI conference on human factors in computing systems. ACM, pp 981–992
5. Chao J, Wang H, Zhou W, Zhang W, Yu Y (2011) Tunesensor: a semantic-driven music recommendation service for digital photo albums. In: Proceedings of the 10th International Semantic Web Conference. ISWC2011
6. Chen HC, Chen AL (2001) A music recommendation system based on music data grouping and user interests. In: Proceedings of the tenth international conference on Information and knowledge management. ACM, pp 231–238

7. Costa YM, Oliveira LS, Silla CN Jr (2017) An evaluation of convolutional neural networks for music classification using spectrograms. Appl Soft Comput 52:28–38
8. Defferrard M, Benzi K, Vandergheynst P, Bresson X (2016) Fma: a dataset for music analysis. arXiv:1612.01840
9. Dias R, Fonseca MJ (2013) Improving music recommendation in session-based collaborative filtering by using temporal context. In: 2013 IEEE 25Th international conference on tools with artificial intelligence. IEEE, pp 783–788
10. Goto M (2006) A chorus section detection method for musical audio signals and its application to a music listening station. IEEE Trans Audio Speech Lang Process 14(5):1783–1794
11. Harte C, Sandler M, Gasser M (2006) Detecting harmonic change in musical audio. In: Proceedings of the 1st ACM workshop on Audio and music computing multimedia. ACM, pp 21–26
12. Hua XS, Lu L, Zhang H (2004) Automatic music video generation based on temporal pattern analysis. In: Proceedings of the 12th annual ACM international conference on Multimedia. ACM, pp 472–475
13. Hua XS, Lu L, Zhang H (2004) Optimization-based automated home video editing system. IEEE Trans Circ Syst Video Technol 14(5):572–583
14. Liao Z, Yu Y, Gong B, Cheng L (2015) Audeosynth: music-driven video montage. ACM Trans Graph (TOG) 34(4):68
15. Lin JC, Wei WL, Wang HM (2016) Automatic music video generation based on emotion-oriented pseudo song prediction and matching. In: Proceedings of the 2016 ACM on Multimedia Conference. ACM, pp 372–376
16. Lin JC, Wei WL, Wang HM (2016) Demv-matchmaker: emotional temporal course representation and deep similarity matching for automatic music video generation. In: 2016 IEEE international conference on Acoustics, speech and signal processing (ICASSP). IEEE, pp 2772–2776
17. Murch W (2001) In the blink of an eye: A perspective on film editing. Silman-James Press
18. Nam J, Tewfik AH (2005) Detection of gradual transitions in video sequences using b-spline interpolation. IEEE Trans Multimed 7(4):667–679
19. Park CW, Young SM (1986) Consumer response to television commercials: The impact of involvement and background music on brand attitude formation. J Market Res 23(1):11–24
20. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
21. Strähle J, Hohls R (2018) In-Store Music in Fashion Stores. Springer, Singapore, pp 71–92. https://doi.org/10.1007/978-981-10-5637-6_5
22. Wang J, Chng E, Xu C, Lu H, Tian Q (2007) Generation of personalized music sports video using multimodal cues. IEEE Trans Multimed 9(3):576–588
23. Wang JC, Yang YH, Jhuo IH, Lin YY, Wang HM et al (2012) The acousticvisual emotion guassians model for automatic generation of music video. In: Proceedings of the 20th ACM international conference on Multimedia. ACM, pp 1379–1380
24. Wang X, Wang Y (2014) Improving content-based and hybrid music recommendation using deep learning. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM, pp 627–636
25. Wu X, Qiao Y, Wang X, Tang X (2012) Bridging music and image: a preliminary study with multiple ranking cca learning. In: Proceedings of ACM Multimedia
26. Yapriady B, Uitdenbogerd AL (2005) Combining demographic data with collaborative filtering for automatic music recommendation. In: International conference on knowledge-based and intelligent information and engineering systems. Springer, pp 201–207
27. Zhang K, Zhang H, Li S, Yang C, Sun L (2018) The pmemo dataset for music emotion recognition. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. ACM, pp 135–142

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Hui Zhang received the B.Eng. degree in computer science and technology from Zhejiang University, China, in 2016. She is currently pursuing the Ph.D. degree in Digital Art and Design with the college of computer science and technology, Zhejiang University, China. Her research interests include music information retrieval, affective computing and humancomputer interaction.



Kejun Zhang received Ph. D. Computer Science from Zhejiang University, China. He is currently an associate professor in the College of Computer Science, Zhejiang University, China. His research interests lies in music information retrieval, artificial intelligence and machine learning. He has published many research papers in various reputable journals and conference proceedings.



Yingping Cao received the Bachelor's degree of Industrial Design, Zhejiang University, Hangzhou, China, in 2017, where she is pursuing the master's degree. Her research interests include music affection, user experience and user shopping behavior research.



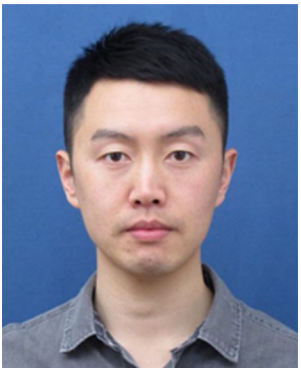
Jun Zheng received the Bachelor of Engineering degree of Computer Science and Technology, Zhejiang University, in 2017, where she is currently pursuing the master's degree. Her research interests include natural language processing and audio processing.



Xiaoyi Huang is currently pursuing the Master's degree of Design from the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. Her research interests include video automatic generation, computing of music structure and interaction design.



Changyuan Yang is a senior expert in Alibaba Group. He is the Director of Human-Computer Natural Interaction Laboratory in Hangzhou. His research interests include design knowledge graph, design intelligence and multimodal interaction methodology.



Lingyun Sun is a professor at Zhejiang University. He is the Deputy Director of International Design Institute, Ng Teng Fong Chaired Professor and the director of ZJU-SUTD Innovation, Design and Entrepreneurship Alliance, and a deputy director of ZJU-Beidou Joint Innovation Design Engineering Center. His research interests include Design Intelligence, Innovation and Design, Information and Interaction Design. He is a PI to research grants funded by National Natural Science Foundation and National Basic Research Program of China. He has more than 30 publications in peer-viewed journals and conferences, include Design Studies and Science China. He is also the inventor of more than 20 granted patents as well as the winner of 5 IF and Red Dots awards.

Affiliations

Hui Zhang[1] · Kejun Zhang[1] · Yingping Cao[1] · Jun Zheng[1] · Xiaoyi Huang[1] · Changyuan Yang[2] · Lingyun Sun[3]

Hui Zhang
huiz@zju.edu.cn

Yingping Cao
elainecao@zju.edu.cn

Jun Zheng
zzjderry@gmail.com

Xiaoyi Huang
choyeehuang@gmail.com

Changyuan Yang
changyuan.yangcy@alibaba-inc.com

Lingyun Sun
sunly@zju.edu.cn

1 Laboratory of CAD, CG, Zhejiang University, Hangzhou, China

2 Alibaba NHCI Lab, Alibaba Group, Hangzhou, China

3 International Design Institute, Zhejiang University, Hangzhou, China