

Automatic Music Video Generation Based on Temporal Pattern Analysis

Xian-Sheng HUA, Lie LU, Hong-Jiang ZHANG

Microsoft Research Asia

{ xshua; llu; hjzhang }@microsoft.com

ABSTRACT

Music video (MV) is a short film meant to present a visual representation of a popular music song. In this paper, we present a system that automatically generates MV-like videos from personal home videos based on observations that generally there are obvious repetitive visual and aural patterns in MVs. Based on a set of video and music analysis algorithms, the automatic music video (AMV) generation system automatically extracts temporal structures of the video and music, as well as repetitive patterns in the music. And then, according to the structure and patterns, a set of highlight segments from the raw home video footage are selected, aiming at matching the visual content with the aural structure and pattern. And last, the output music video is rendered by connecting the selected highlight video segments with appropriate transition effects, accompanied with the music. Experiments show that the results are compelling and promising.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems — video; I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—video analysis.

General Terms

Algorithms, Experimentation.

Keywords

Video content analysis, video segmentation, music analysis, music video, video editing, optimization.

1. INTRODUCTION

Music video (MV) originally started from the year of 1950. It is a short film meant to present a visual representation of a popular music song. Typically MVs are recorded using top film equipments or professional video cameras, which may cost top to 25 to 10 thousands dollars each [1]. While camcorder becomes a commodity home appliance, common users have desires to produce their own MVs using non-professional tools. Existing video editing systems, such as *Abode Premiere*, are a great help for MV creating, but the task is still a tedious and time consuming, requiring significant editing skills and an aesthetic sense. Therefore an automatic tool is demanded. In this paper, we present a system that automatically generates MV-like videos using prerecorded personal home videos and user provided music, based on a set of video and music analysis algorithms.

To generate an enjoyable MV, the typical characteristics of MVs are necessary to be analyzed. It is observed that typically there are

obvious repetitive patterns in MVs, both in video tracks and audio tracks. For example, a song generally contains several repetitive sections and other instrumental sections such as *prelude*, *interlude*, and *coda*. Usually the video track has also a corresponding property as audio track. That is, typically the visual content in prelude, interlude and coda are similar, so as the visual content for the repetitive sections (see Figure 2, the corresponding visual segments of the four occurrences of the repetitive music pattern MP_1 are most likely similar). In addition, unlike professional MV production, in which the video are taken with clear intention, to generate MVs from raw home video footage, how to select appropriate segments from them is also critical due to the visual quality of home videos is generally very low, as well as there are many redundant or less attractive segments.

Previous works on automatic MV production are reported in [2] and [3] in the literature. In [2], video segment is selected based on video unsuitability, and then merged together along the music timeline without taking music repetitive pattern and motion-tempo matching into consideration. In [3], an Automated Home Video Editing system is proposed, in which MV generation is one of the four “editing styles”. For this style, the durations of music sub-clips are determined by the average tempo of the music. That is, a fast music clip will result in fast shot changes in the output video, and vice versa. And, the motion intensities of the video segments and tempos of the music clips are well matched in that work. However, music repetitive pattern is not taken into account either.

In this paper, an automatic MV (AMV) generation system is proposed to automatically analyze the temporal structure of the raw home videos and the user provided music, as well as the repetitive patterns in the music. Then, according to the temporal structure and patterns, a set of highlight segments from the raw home video footage are selected, in order to appropriately match the visual content with the aural structure and repetitive patterns.

Figure 1 illustrates the work flow of the proposed AMV system, which consists of four major steps, including Preliminary Media Content Analysis, Video Scene – Music Pattern Matching, Video Repetitive Pattern Generation, and Final Rendering.

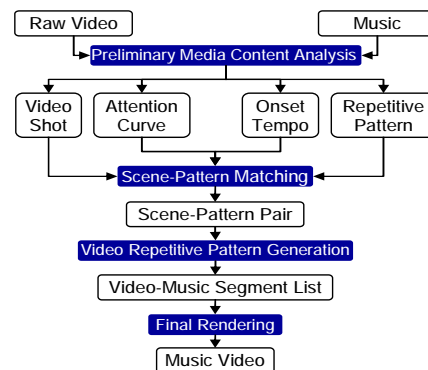


Figure 1. Flow chart of Automatic Music Video Generation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'04, October 10–16, 2004, New York, NY, USA.

Copyright 2004 ACM 1-58113-893-8/04/0010...\$5.00.

The rest of the paper is organized as follow. Section 2 introduces preliminary content analyses both for video and music. The rest three steps of AMV are detailed in Section 3, followed by experiments in Section 4, and conclusion remarks in Section 5.

2. CONTENT ANALYSIS

To automatically generate MVs, a set of content-based features of the raw home video and user selected music are extracted, which are the basis of the proposed AMV system.

2.1 Video Analysis

With video analysis, the raw home video is segmented into shots according to color similarities or timestamp (if it is provided or able to be recognized); and then an “attention” or “importance” value of each shot is calculated by averaging the “attention index” of each video frame, in which the attention index is the output of “attention detection” [5] relating to object motion, camera motion, color and speech in the video. Suppose the video is represented by a shot series as $V = \{Shot_i, 0 \leq i < N\}$, where N is the number of shots, the “importance” of each shot is denoted by $I(Shot_i)$.

Based on the above analysis, the following information is also obtained (similar to [3][5]): camera motion (type and speed), motion intensity (denoted by $MI(\cdot)$) and color entropy, which will be employed in later steps.

2.2 Music Analysis

2.2.1 Onset Detection and Rhythm Estimation

Onset is the moment when a key is pressed down, which is used to roughly estimate the music rhythm in our system, as well as align music clip and video shot boundaries. An onset detector proposed in our previous work [3] is applied to extract the onset series, as well as the corresponding onset strengths in the incidental music by checking “energy peaks” in frequency domain. The onset sequence is finally denoted as $Onset = \{(T_i, S_i), 0 \leq i < L\}$, where T_i is the time of the i -th onset, S_i is the corresponding strength, and L is the number of onsets in the music.

The tempos of the entire music and any sub-clip (i.e., a music segment) are roughly estimated as the onset frequency in it, denoted by $Tempo(m)$, where m is the music or a music sub-clip. The higher the value is, the faster the tempo is.

2.2.2 Repetitive Pattern Discovery

Music generally shows strong self-similarity, and thus has some repeating patterns and prominently repetitive structure. These repeating patterns and structure are very helpful for MV generation. In this section, we adopt a new approach to extract all the significant repetitions that have similar melody as reported in [4]. First, each feature set is extracted from the acoustic data, including temporal feature, spectral feature and CQT feature. Temporal features are used to estimate tempo period and the length of a musical phrase, which is used as the minimum length of a significant repetition in repeating patterns discovery and boundary determination. Spectral features are used for vocal and instrumental sounds discrimination in order to identify the *prelude*, *interlude* and *coda* of a popular song in final music structure analysis. CQT features are used to represent the note and melody information, based on which a self-similarity matrix of the music is obtained. The significant repeating patterns are then detected from the similarity matrix with an adaptive threshold setting method. Finally, the boundaries of repeating patterns are roughly aligned to facilitate music structure inference; and the obtained structure is utilized correspondingly to refine the

boundary of each musical section, with an optimization-based approach.

The output of music repetitive pattern discovery is represented by

$$MP_i = (Type_i, Num_i, \{MS_{ij} = (Start_{ij}, End_{ij}, Tempo_{ij}), 0 \leq j < Num_i\}) \quad (1)$$

where $0 \leq i < K$, K is the number of patterns (*prelude*, *interlude*, and *coda* are regarded as one pattern called “instrument”); $Type_i$ and Num_i are the type (*normal* or *instrument*, the former one covers all non-instrument patterns); $(Start_{ij}, End_{ij}, Tempo_{ij})$ are the start time, end time and tempo of j -th occurrence of a music pattern MP_i in the music. Except “instrument” pattern, generally the tempos of different occurrences of the same music pattern will be the same. Figure 2 shows a typical example of music pattern detection results, which consisting of one *instrument* (denoted by MP_0 in the figure) and two normal repetitive patterns (denoted by MP_1 and MP_2 , respectively), while MP_1 and MP_2 have four and two occurrences respectively.

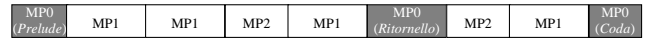


Figure 2. Example of Music Pattern.

3. MUSIC VIDEO GENERATION

In this section, we present the rest three major steps of AMV after content analysis which has been introduced in previous section.

3.1 Video Scene - Music Pattern Matching

This step aims at finding an appropriate set of shots (i.e., scene) for each music repetitive pattern from video V under certain constraints. It consists of three sub-steps, including quality filtering, scene segmentation, and scene – music pattern matching.

Firstly the low-quality shots are filtered out using similar method as [3] which is based on camera motion speed and color entropy (to be exact, shots with very low color entropy or extremely high camera motion speed are removed from the shot list). For convenience, we still use $V = \{Shot_i, 0 \leq i < N\}$ to denote the shot list after quality filtering.

Then the shots are grouped into αK to βK scenes according to content similarity and timestamp (if available), where α and β are two parameters that are experientially set to 1.5 and 3 in current implementation (recall K is the number of musical repeating patterns). In particular, we define the similarity of any two consecutive shots (or we may say it is the similarity measure of the “connection point” of these two shots) as the weighted sum of a series of histogram intersections of the shots at the two sides of the “connection point”:

$$Sim_i = Sim(Shot_i, Shot_{i+1}) = \sum_{j=1}^{\delta} \beta_j S_{i,i+j} + \sum_{j=-\delta}^{-1} \beta_{-j} S_{i+j+1,i+1} \quad (2)$$

where $0 \leq i < N-1$. $S_{k,l}$ is the histogram intersection (i.e., color similarity) of $Shot_k$ and $Shot_l$ in HSV color space, and the parameter β_j is the summing weight defined by

$$\beta_j = \beta^j, 0 < \beta \leq \delta \quad (3)$$

in this implementation ($\beta = 2/3$, $\delta = 5$). If timestamp if provided, we may use a method similar to the one in [6], which originally was used for photo clustering, to obtain better results.

Figure 3 shows an example of shot similarity curve Sim_i . Scene segmentation is equivalent to finding a set of cut points in this curve. Assume we will cut the shot list at point list θ , while θ is an $(\alpha K-1)$ -element to $(\beta K-1)$ -element subset of $\{0, 1, \dots, N-2\}$, and Θ

the set of all subsets of this form. Points in curve Sim_i whose subscripts are in θ are the cut points. Then, we grouped the shots into αK to βK groups by solve the below optimization problem:

$$\theta^* = \arg \min_{\theta \in \Theta} \frac{\sum_{j \in \theta} Sim_j}{|\theta|} \quad (4)$$

where $|\theta|$ stands for the number of elements in the finite set θ . *Genetic Algorithm* (GA) [7] is applied to find global optimal solution for this optimization problem, as well as for the other ones to be presented in this paper later.

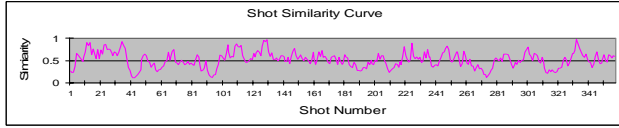


Figure 3. Example of Shot Similarity Curve.

Suppose finally the shots are grouped into K^* scenes, represented by $Scene = \{Scene_i, 0 \leq i < K^*\}$, using the above algorithm. Then next task is to select K scenes from them which correspond to the K music patterns. To make the final music video compelling and more like a professional edited music video, we try to match the tempos of the music repetitive patterns with the motions intensities in the corresponding video scenes, as well as to preserve the “important” segments in the raw video. To be exact, our objective is to simultaneously maximize the average scene importance and the *Correlation Coefficient* between the tempo sequence (of the music patterns) and the motion intensity sequence (of the selected scene series). Similar to scene segmentation, suppose φ is a K -element subset of $\{0, 1, \dots, K^*-1\}$, and the elements in $Scene$ whose subscripts are in φ are the scenes will be selected, while Φ the set of all subsets of this form. Accordingly, an optimal φ is determined by finding φ^* as

$$\varphi^* = \arg \max_{\varphi \in \Phi} \left(\gamma_1 \frac{\sum_{j \in \varphi} I'(Scene_j)}{|\varphi|} + \gamma_2 \rho(\varphi) \right) \quad (5)$$

where $\gamma_1 + \gamma_2 = 1$ are weights for the combination (both are set to 0.5 here), and $\rho(\varphi)$ is the correlation coefficient of tempo sequence $\{Tempo(MP_i), 0 \leq i < K\}$ and motion intensity sequence $\{MI(Scene_j), j \in \varphi\}$. $I'(Scene_j)$ is defined by

$$I'(Scene_j) = \frac{\sum_{Shot_m \in Scene_j} I(Shot_m)}{\sqrt{|Scene_j|}} \quad (6)$$

Here we average the “importance” by square root of $|Scene_j|$ instead of $|Scene_j|$ since we also take the size (number of shots) of the scene into account to a certain extent. That is, the larger the size is, the more likely the scene is “important”.

3.2 Video Repetitive Pattern Generation

In previous step, each music pattern is assigned with a video scene. In this step, from this assigned scene, repetitive video segments associated with the corresponding repetitive music segments are extracted, thus the repetitive patterns in audio track and video track are accordant.

For a certain music pattern MP_i defined as Equation (1), suppose the corresponding assigned scene is $Scene_i = \{Shot_{i,m}, 0 \leq m < M_i\}$, where M_i is the number of shots in the scene. In order to select appropriate video segments for the music segments, as well as align shot transitions with the strong onsets in music, firstly we divide every occurrence of the specific music pattern MS_{ij} (it is a

music segment indeed) into small music sub-clips by finding strong onsets in a sliding window as following steps:

(1) To make the shot change frequency in the final video is accordant with the tempo (recall it is estimated by onset frequency), the size of the sliding window is adaptively determined by the value of the estimated tempo. That is, we restrict the minimum (Min_{ij}) and maximum (Max_{ij}) duration of the music sub-clips by

$$\begin{aligned} Min_{ij} &= \min(\max(Min, 1/Tempo_{ij}), Max-1) \\ Max_{ij} &= \min(Max, Min_{ij}+1) \end{aligned} \quad (7)$$

where Min and Max are predefined threshold (set to 1 and 5 here, in seconds), respectively. And “min” and “max” are functions to get the minimum and maximum elements, respectively. Accordingly, the duration of music sub-clip is in inverse proportion to the music tempo.

(2) The start time of MS_{ij} , $Start_{ij}$, is taken as the first boundary, denoted by B_l , while $l = 0$;

(3) Find the strongest onset in time window $[B_l + Min_{ij}, B_l + Max_{ij}]$, and assign the time of the strongest onset to B_{l+1} , and then set $l = l+1$.

(4) Continue (3) until all boundaries are found.

Suppose there are L_{ij} music sub-clips in segment MS_{ij} , denoted by $SC_{ij} = \{SC_{ij,l}, 0 \leq l < L_{ij}\}$. As has presented, to keep consistency of the visual content for a certain music pattern, the corresponding video segments are selected from the same scene. While to keep the variation of visual content along the timeline for a certain music repetitive pattern, the corresponding video segments for different occurrences of the certain music pattern are selected from different sub-shots (a segment within a shot, see [3]) of the shots in the assigned scene. Here we assume that $L_{ij} \leq M_j$, so we can select L_{ij} shots from $Scene_i$, thus we are able to assign one different shot for each music sub-clip. If this constraint is not satisfied, which may occur occasionally, we can either repeat some important shots, or split long shots till this constraint is met.

Similar to scene selection in Section 3.1, we will also find an optimal set of shots from the $Scene_j$ for the music sub-clips. Suppose ζ is a L_{ij} -element subset of $\{0, 1, \dots, M_j - 1\}$, and the elements in $Scene$ whose subscripts are in ζ are the shots will be selected, while H the set of all subsets of this form. Accordingly, an optimal ζ is determined by finding ζ^* as

$$\zeta^* = \arg \max_{\zeta \in H} \left(\lambda_1 \frac{\sum_{m \in \zeta} I(Shot_{im})}{|\zeta|} + \lambda_2 \rho(\zeta) \right) \quad (8)$$

where $\lambda_1 + \lambda_2 = 1$ are combination weights (set to 0.5 here), and $\rho(\zeta)$ is the correlation coefficient of tempo sequence $\{Tempo(SC_{ijl}), 0 \leq l < L_{ij}\}$ and motion intensity sequence $\{MI(Shot_{im}), m \in \zeta\}$.

Suppose the optimal set of shots from $Scene_i$ for MS_{ij} is $\{Shot_{i,l}^{(l)}, 0 \leq l < L_{ij}\}$, and the start time and duration of a video or music segment x is represented by $S(x)$ and $D(x)$, respectively, then the corresponding video sub-shot for music sub-clip SC_{ijl} is determined by $SS_{ijl} = (Start_{ijl}, End_{ijl})$ as

$$\begin{aligned} Start_{ijl} &= S(Shot_{i,l}^{(l)}) + (D(Shot_{i,l}^{(l)}) - D(SC_{ijl})) \cdot l / (L_{ij} - 1) \\ End_{ijl} &= Start_{ijl} + D(SC_{ijl}) \end{aligned} \quad (9)$$

That is, different portions of the same shot will be applied for different occurrence of the same music pattern in the same position, thus the corresponding video segments for different occurrence of a certain music pattern will have relatively high similarity (as they belong to the same set of shots) while not completely the same.

Consequently, a video-music sub-clip list, (SC_{ijl}, SS_{ijl}) , where $0 \leq i < K$, $0 \leq j < Num_i$, $0 \leq l < L_{ij}$, is obtained, from which the final music video will be created, as to be presented in next sub-section.

3.3 Final Rendering

Generally we can obtain the music video by connecting the sub-clip list (SC_{ijl}, SS_{ijl}) after sorting by the start time of all the music clips. However, to emphasize the repetitive patterns, we adopt the following rules based on aforementioned observations.

- (1) *Transition Rule*: According to observations on typical MVs, cut and cross-fade are most commonly used transitions. In AMV, transitions connecting different music patterns are always “slow” cross-fade (i.e., transition duration will be about 1 second). For other transitions, if the two consecutive sub-shots are similar in color (refer to Section 3.1), then use cross-fade and the transition duration is determined by the strength (S) of the onset on the connecting point as

$$TransitionDuration = \max(0.25, 1 - S). \quad (10)$$

Otherwise, cut is applied (transition duration is zero for cut).

- (2) *Effect Rule*: For the sub-shots associated with *prelude*, *interlude* and *coda* are applied with certain transformation effects, such as *Sepia Tone*, *Grayscale*, *Slow Motion* or *Old Movie* [3], to enhance the feeling of pattern repetition.
- (3) *Start/End Sub-Shot Rule*: The last sub-shot of the music video is taken from the same shot where the first sub-shot taken from. And, the first sub-shot will be applied with a “*Fade In*” effect, and the last sub-shot will be applied with a “*Fade Out*” effect. In addition, captions (music name, singer’s name, etc.) of the music will be superimposed on the bottom-left corner of the first sub-shot and the last sub-shot.

4. EXPERIMENTS

Figure 4 shows the process of generating a 3-minute MV from a 90-minute raw home video. Figure 4 (a) is the results of scene grouping, where the selected scenes are also indicated in it. The corresponding music patterns are the ones illustrated in Figure 2. Figure 4 (b) shows the relations between the scenes and music patterns. Due to space limitation, the shot and sub-shot level matching results are not listed here.

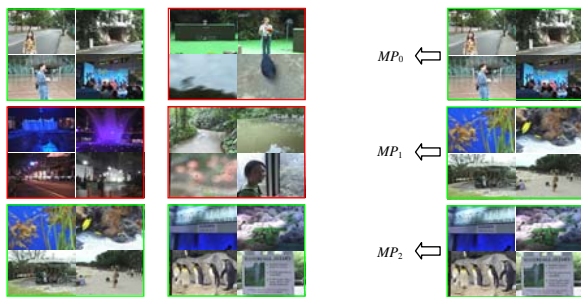


Figure 4. Example of AMV generation.

As it is difficult to objectively evaluate AMV system, we compare AMV with other two related systems by subjective evaluation: one is AVE - MV style, and the other is video summarization (VS), i.e., connecting the most “important” segments without taken video-music matching in to account, while music is still accompanied. Five pairs of raw home videos (of various content including vacation, scenery, wedding, etc.) and music (of various genres including pop music, classical music, etc.) are applied in the user study. Therefore there are 15 music videos in total, which are provided to ten users who are required to give a satisfaction

score between 0 and 1 to each of the 15 music videos. The videos are randomly ordered and the users do not know which video is generated by which method. Table 1 lists the average evaluations, which indicate that AMV results are generally better than AVE, while AVE results are better than VS results. Several sample videos are available at <http://research.microsoft.com/~xshua/amv/> for download. Due to lack of resource, as well as our objective is not to provide a professional music video generation scheme but an automatic tool to get MV-like videos thus the users may continue working on these results, we have not compared AMV with professional editing results (though it is our future work).

Table 1. Subjective evaluation and comparison.

#	Video		Music		Avg Score		
	Content	Duration	Genre	Duration	VS	AVE	AMV
1	Vacation	1 h 30 m	Pop Song	3 m 5 s	0.67	0.76	0.80
2	Scenery	20 m	Classical	2 m 30 s	0.60	0.69	0.75
3	Wedding	2 h 10 m	Light	4 m 25 s	0.70	0.79	0.77
4	Vacation	1 h 5 m	Pop Song	3 m 10 s	0.72	0.82	0.85
5	Xmas	40 m	Light	2 m 5 s	0.58	0.65	0.88
Avg	-	-	-	-	0.65	0.74	0.81

5. CONCLUSION AND FUTURE WORK

In this paper, we present a system that automatically generates MV-like videos from personal home videos based on observations that generally there are obvious and accordant repetitive visual and aural patterns in typical MVs. Based on a set of video and music analysis algorithms, the AMV system automatically extracts temporal structures and/or repetitive patterns in the video and music. And then, according to the structure and patterns, a set of highlight segments from the raw home video footage are selected, which are matched with the aural structure and patterns of the music. And last, the output music video is formed by connecting the selected video segments with transition effects, accompanied with the music.

Obviously, we may generate music video in other ways. This paper only provides one of many possible solutions or styles of automatic music video generation. For example, another solution would be to preserve a storyline of the raw home video, while add several repetitive patterns among them according to the music repetitive patterns. This is one of our future works. Furthermore, AMV can also work in an interactive manner, in which the system provides suggestions and matches for users or artists to create MVs. In addition, we are also planning to investigate the shot connecting rules or patterns for typical music videos, as well as integrating face detection results and video semantic classification technologies, to make the output videos more compelling and more like professional MVs.

6. REFERENCES

- [1] Dwelle, T. Music Video 101. [eBook] <http://www.timtv.com>.
- [2] Foote, J., et al. Creating Music Videos Using Automatic Media Analysis. *ACM Multimedia 2002*.
- [3] Hua, X. S., et al. Optimization-Based Automated Home Video Editing System. *IEEE Trans. on Circuits and Systems for Video Technology*. Vol 14, No. 5, May 2004, 572-583.
- [4] Wang, M., et al. Repeating Pattern Discovery from Acoustic Musical Signals. *Intl Conf. on Multimedia and Expo, 2004*.
- [5] Ma, Y. F., et al. A User Attention Model for Video Summarization. *ACM Multimedia 2002*, 533-542.
- [6] Platt, J. AutoAlbum: Clustering Digital Photographs using Probabilistic Model Merging. *IEEE Workshop on Content-Based Access to Image and Video Libraries, 2000*.
- [7] Whitley, D. A Genetic Algorithm Tutorial. *Statistics and Computing*, Vol. 4, 64-85, 1994.