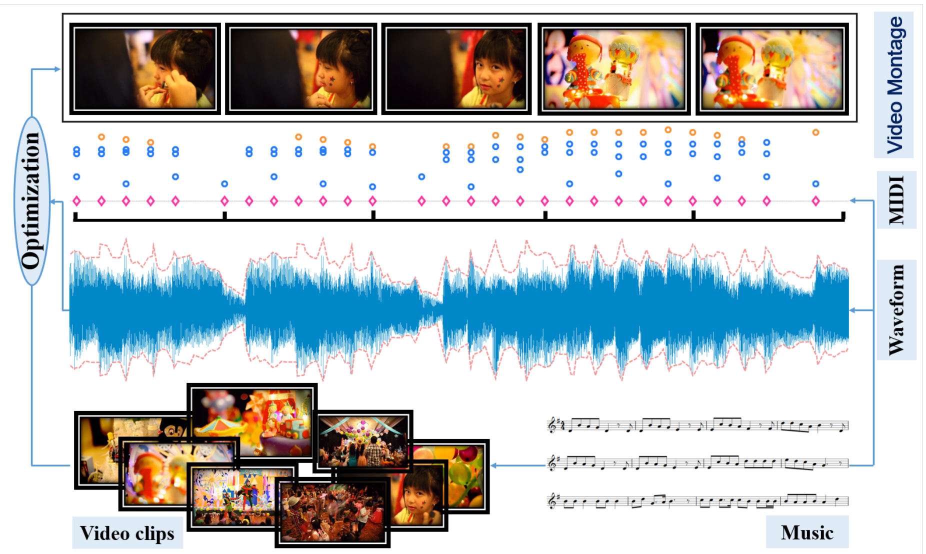
PAPER ANALYSIS

TOPIC:
audeosynth: Music-Driven Video Montage

TANYA KHEMANI
khemanit@oregonstate.edu
CS 550

audeosynth: Music-Driven Video Montage



The paper “audeosynth: Music-Driven Video Montage” introduces a video montage which is driven by music. It facilitates the user to browse the videos clips collected from different occasions. The title name itself suggests that the music drives the composition of the video content. “Montage” describes a sequence that uses a collection of short clips to show a passage of time. Music and videos are two different types of media. In this paper, they create a framework for generating the music-driven video montages automatically. This step was taken to overcome the problem of applying enormous manual work and artistic expertise for the videos organized to form a montage that reflects the properties of the music. So to get over with this problem, a system driven approach is designed, where the system takes the input as a set of video clips and synchronizes it with some background music. This technique is tough because visual activities are synchronized with the rhythm of the music. As a result, the synthesis problem is solved through Markov Chain Monte Carlo sampling which gives a video montage as the output from which short clips are taken and synchronized with the rhythm of music to generate an audio-video resonance.

The authors Zicheng Liao, Yizhou Yu, Bingchen Gong and Lechao Cheng are from Zhejiang University and The University of Hong Kong.

Zicheng Liao is an associate professor, Ph.D and now works at the College of Computer Science, Zhejiang University in Hangzhou, China. He did his Bachelors from the same University. He received his Ph.D. degree in Computer Science from University of Illinois at Urbana-Champaign (UIUC) under the advice of David Forsyth. His research works focus on Computer Vision, Machine Learning and Computer Graphics. Talking about his industrial experience, he was a Microsoft Research Intern with Hugues Hoppe of the Graphics group for Summer 2012 and 2013 in Redmond. He worked as a Data Scientist intern at Facebook for Summer 2009 and 2010 in Palo Alto. Also, he was a Software Engineer Intern for Summer 2007 in Beijing.

Yizhou Yu is currently a full professor in the Department of Computer Science at the University of Hong Kong, and an adjunct professor in the Department of Computer Science at University of Illinois, Urbana-Champaign (UIUC). He was first a tenure-track and then a tenured professor at UIUC for more than 10 years. He has also collaborated with eBay Research, Google Brain and Microsoft Research in the past. He received his PhD degree in computer science from the computer vision group at University of California, Berkely. He also holds a MS degree in applied mathematics and a BE degree in computer science and engineering from Zhejiang University. Prof Yu has made many important contributions to AI and visual computing, including deep learning, computer vision, image processing, graphics, and VR/AR. He is a recipient of 2002 US National Science Foundation CAREER Award, 2007 NNSF China Overseas Distinguished Young Investigator Award, 2011 and 2005 ACM SCA Best Paper Awards, and 1998 Microsoft Graduate Fellowship. Innovative technologies co-invented by him has been frequently adopted by the film and special effects industry. He has more than 100 publications in international conferences and journals. His current research interests include deep learning methods for machine intelligence, computational visual media, geometric computing, intelligent video surveillance, and biomedical data analysis.

Bingchen Gong is currently a student in the department of Computer Science and Technolgy at Zhejiang University. His expertise and skills include Machine Learning and Computer Vision.

Lechao Cheng is currently a Ph.D student at Zhejiang University in the College of Computer Science and Technology. His current research centers around computer vision and deep learning.

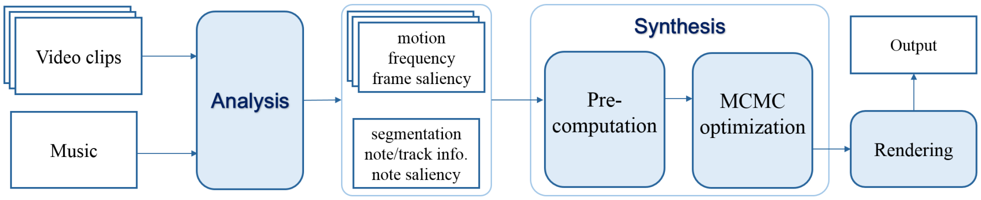
The authors take various ideas from music driven imagery. They state that a much better way of browsing video clips is with a music driven montage. As the main goal of this paper is to create a computer aided solution for creating an audio-visual composition, the authors suggest synchronization between music and video by the method of matching cost. In this method, the ups and downs of a video sequence is matched with that of the music. However, as music and video are two different media, which becomes a drawback for the approach. As a result, the authors introduce a duplication cost penalty for this approach, which is a positive thing since efficiency matters a lot for the applications running on mobile devices and also an additional cost is saved as some segments of video and music may be reused as the penalty is introduced.

This application can run very well on mobile devices such as cell phones as the users capture a lot of videos in their day to day life and browse them during their spare time. Creating a high quality music video and integrating it with an algorithm is a very laborious task. Therefore, the authors also discuss about some of the following factors that may make the software more difficult:

1) The selection of the content of the video to be displayed. Because not all the frames of the video show the accurate scenes that can be included, then that should be detected by the software.
2) The play speed of the videos as some of the particular videos require slow motion. This generally happens when the video shows some appealing scene. Therefore, the software is also responsible for detecting this.

3) Another important factor found by the authors is where to apply the cut in a particular video. As a result, they adopt two thumb rules such as cut-to-the-beat and synchronization.

In order to solve the serious optimization problems, the authors suggested applying Markov Chain Monte-Carlo sampling. The following three stages were suggested:
1. Analysis Stage
2. Synthesis Stage
3. Rendering



Given a set of video clips and a piece of music, our system first performs feature analysis from both inputs, and then cast the synthesis problem as an optimization problem, solved by a two-stage algorithm. Lastly, the rendering component renders the final video montage that are cut and synchronized to the beat of the music.

The way in which these stages work is the analysis stage performs signal analysis on video clips and the music piece and then sends it to the synthesis stage. Some of the important features from the video frames are extracted such as motion analysis, saliency estimation and frequency detection. For music analysis, information is collected from the MIDI data. Features such as note onsets, pitch, duration are taken into consideration while performing music analysis. Then bars are segmented into music segment on top of features such as pace, volume, pitch variation, number of tracks. They kept the synthesis stage to get input from the analysis stage. Both the features of music and video are taken by the synthesis stage. The rendering stage states the mapping of the video segment to the audio segment. It guarantees that music won't be cut. This section mostly focus on mapping.

The video analysis only measures the visual activities in a video using optical flows. It does not measure the deep art in those activities. Such system may mistake a sarcastic video for a happy or a sad video. The current system does not blend in the human emotions. This can be a major disadvantage of this system. Object movement is not enough to measure the emotion of a particular video. In the part of motion analysis, it is explained that the most important motion might be a baseball player hitting a ball. Therefore, they state that when the ball hits the bat, and goes somewhere, is the important moment. But this is not enough. Consider a clip where a baseball player is hitting many balls. Such system will fail to detect the important ball. According to the algorithm, it will detect all of the balls. But it will fail to identify the most important ball. Consider an instance of clip which consists of World Series Final and the game is tied to the last inning, will the algorithm detect the final important ball as most significant than the ball from the ball from the first inning. The answer is no, this is because, the algorithm does not dwell in finding most significant events, rather than significant events in common tasks. Therefore, this aspect can make it effective, but it won't be most effective.

For calculating saliency, the authors propose eight types of saliency scores. They prefer keeping these at zero in the start. The saliency scores will turn to one when certain conditions are met. The disadvantage of this approach is that if a certain media is said to be sad then it will have a pitch shift. But the binary saliency won't be turned one in the start and the media will have to be rerun in order to do this. This also represents wastage of some resources. Therefore, this technique can be made more effective by a certain approach. The paper suggests a condition in which two types of saliency matches. Although, it does not suggest when more than two types of saliency matches. This proposes a disadvantage to this approach. Because some of the songs may have more than three types. In this approach, some of the saliency scores seem to delineate the other scores. Therefore, this technique may not prove to be always effective. Due to this limitation, the final saliency score is calculated may not be accurate. The global cost of the function may also seem to change due to this. If there is a prefetching in this concept of saliency, then the resources can be saved at large.

The authors suggest two stage optimization for minimizing the energy functions. The Metropolis Hasting algorithm is used for the process of optimization. This function has

disadvantage of not performing very well in large performing space. Some other method could have been preferred for this application. Due to this method, the application cannot work on large media. If they find some other traversing method good for bug space, then this algorithm can be much more effective. The audio-video synchronization acts like a big advantage for this application. This is because of the scalable sliding window, which searches for video sequences that aligns with music segments. This phenomenon is effective due to the concept of dynamic programming. But what if a particular video sequence is matched by an opposite type of music? The dynamic programming can fail if the mechanics to detect cannot differentiate between sarcastic happiness and sadness. Therefore, this technique can work on most of the music. In future, we never know how the media will evolve. This application also has certain disadvantages in the future due to the MIDI format. This technique can be deemed unsuitable in the near future.
In order to test this feature, a user study was done by the authors. They categorized the features in four groups. Group A (cut to beat turned off), Group B (synchronization feature turned off), Groups C and D (done by human users). The authors claim that their results were significantly better than the groups C and D. For this study, a total of 29 participants were collected. Each of those participants were asked to view a particular set of 6 videos. There were five results for each set. The participants were given a privilege of watching all the videos at once or one at a time. The scores were 1 for the worst and 5 for the best rating. The user group C is said to be experts group. This is because, they recruited a television and broadcast journalist for it. As a result, it forms a 6 X 5 video set in the user study.

Using MIDI format is one of the limitations of the algorithm. It is however an ageing format. Currently waveform and the MP3 format are burgeoning. Therefore, due to the use of MIDI format, this technique cannot be used in the near future. Or if it has to be used then it has to be modified to work with mp3 and waveform. But implementing new techniques requires lot of work in this field. The authors also describe some of the advantages of the MIDI format over waveform and mp3 like editable audio representation. They suggest that such organization is required to perform deep analysis of the music. Such analysis may also pave way for new methodologies for audio encoding, audio synthesis and audio editing.

The results have been successfully applied to the algorithm to a variety of video clips and background musical pieces. The following are some examples:

1) Aurora: This example comprises a set of aurora and lightening scenes with the background music excerpted from Easy Going by Bjorn Lynne. The video collection has 36 candidate clips, each of which is 3-5 seconds long. The background music is the segment from second 10 to 30 of the original piece. 10 out of the 36 candidate video clips were chosen to generate the montage. This result demonstrated the synchronization of motion change peaks and music beats.

2) City Time lapse: This example talks about time lapse videos shot by professional photographers. Most of the videos have dynamic and rhythmic time lapse motions. This example demonstrates striking effects by cutting and synchronizing time lapse videos with the beat of a piece of music.

3) Happy Birthday: This example demonstrated the application of the algorithm in organizing video collections of gatherings and events to a specific song. The “Happy Birthday to You” song in this example has 6 bars. The 18 input video clips were taken from a birthday party, each 1-7 seconds long. The algorithm was able to find short moments in the videos to match the beat of the music from a pool of random human/artificial activities that were not intended for the rhythm of the song at all, such as the clown show, kids play, and flashing lights.

4) Adventure: This example demonstrates the application of the algorithm in organizing video collections of outdoor activities, such as sky diving, water skiing, and canoeing. The background music was excerpted from Lottyr, Lady of the Hells. Aside from audio-visual synchronization in-between cuts, this example also demonstrates our algorithm’s capability of automatically varying videos play speed to match the pace of the background music. The associated video segments deliver a visual experience that resonates with the music.

5) Wild: This example covers animal, flower, desert and forest scenes in the wild. The background music was excerpted from Exploration. The movements of the animals (the baboon, elephant and penguins), being irregular at first, were successfully temporally scaled and snapped to the musical beats as well.

The optimization part has been implemented in this function solely for minimization of energy. As described earlier, the application is going to be used for mobile devices. Therefore, efficiency is one of the main purposes for this application. This is also because, the mobile devices have limited power. The users expect mobile applications to be more and more efficient with respect to their battery power. Rendering becomes the final stage in which all the video segments are put together with music in a particular sequence. The final product of this is the video-audio synthesis.

Related work developed in this field is sound computation. The most important topic in this is the sound based animation and music driven photograph slideshow. The particular approach taken by the authors exhibit rhythm analysis, which is adopted in dance to music character animation. The concept of audio-visual synthesis also plays some role in the work of the authors. This is because, the authors try a similar approach on video clips. I think the approach taken by the authors is more difficult since the manipulation is passive. There has been a recent work on video editing, which uses active mode. The application of the authors uses passive mode. Therefore, it is more difficult.

The synthesis stage is explained in detail in the paper. But if it gives an example with its working, then it would be easier to understand.

The authors have mentioned that users in a loop will be good advantage for this application. They advise the use of more sophisticated music effects that can be added. They also suggest that a lot of automatic algorithms are available online that search for music and video clips. But they do not discuss the complexity of implementing such approaches.

Music segmentation and onset detection are one of the few major challenges for the application due to use of the MP3. If I would have been the contributor, I would try to replace this MIDI format with MP3 or some more desirable format which allows accurate semantic analysis. The application is able to create automated montage of the video with music. But I feel that, it will have some disadvantage in detection of emotions.

In my view, I think you need users at some point of analyzing in this application. This is because, analyzing algorithms, however good, cannot detect the sarcastic emotions. They can misjudge the emotions. Therefore, it’s good if users help in this approach. With this, the algorithm will be more effective. Since this application will be used on a mobile device, it could consume minimum battery life for maximum performance. As a result, it will be more efficient. However, development of an application that creates the video montage dances in synchronization to the beat of music, is an innovative approach.