The following paper introduces a video montage which is driven by music. The user could browse his videos from various occasions with a music driven montage of it. The videos will be organized in such a way that the music movement and beats will be synced to it. But this syncing of beats will take a enormous effort of manual work and time. It is also a system driven approach, where the system takes the video input and puts it with an audio set. This technique is tough because visual activities are synchronized with the rhythm of the music. The authors Zicheng Liao,Yizhou Yu, Bingchen Gong and Lechao Cheng are from different Universities from China and Hong Kong. The paper suggests that this concept will give a touch of sensation. The authors take various ideas from music driven imagery. They state that browsing video clips with a music driven montage is more of a pleasant way of browsing video clips. The main goal of this paper id to create a computer aided solution for creating such audio-visual composition.

The authors suggest that this type of application could run very well on mobile devices such as cell phones. This is because, users capture videos from their day to day life and browse them during their spare time. Such content would be wonderful to view online and would be able to be share with other people. Creation of a high quality music video is a tedious job. Integrating this process in an algorithm will be a hard job. Since such skill is artistic and the paper does not tell very well how to do this. The authors also discuss other factors that may make making this software more difficult such as the selection of the content of the video to be displayed. This is because, not all of the content will be appealing for the content that is intended to be created. Because not all of the frames of the video show the important scenes. This should also be detected by the software. The authors also talk about the play speed. This is because, some particular videos require slow motion. This happens when the video shows some appealing scene. Therefore, the software should also detect this. Where to apply the cut in a particular video is also an important issue found by the authors. Therefore, they adopt two thumb rules such as cut- to-the-beat and synchronization. Because the issues discussed above are a serious combinational problems.

In order to solve the serious optimization problems, the authors suggest applying Markov Chain Monte-Carlo. Three stages are suggested in the framework i.e. 1.Analysis Stage, 2.Synthresis Stage 3.Rendering stage. The way in which these stages work is that 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 have 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 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. This is also because of the jargon used in this paper. The rendering stage of the paper states the mapping of the video segment to the audio segment. They guarantee that the music won’t be cut. They mostly focus on mapping in this section.

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 wont 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 wont be turned one in the start and the media will have to be re run 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. But 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. This is due to certain conditions. 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 function. The Metroplis Hasting function 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 segemnts. This phenomenon is effective due to the concept of dynamic programming. But what if a particular video sequence is matched by a opposite type of music? The dynamic programming can fail if the mechanics to detect cannot differenciate between sarcastic happiness and sadness. Therefore, this technique can work on most of the music by not on some of the material in the present. 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. Therefore due to those 4 control groups and their result, it forms a 6 X 5 video set in the user study.