MUSIC SYNTHESIS FOR HOME VIDEOS

MEERA G NAYAK

(B.E Electronics and Communication, Bangalore University, India)

A THESIS SUBMITTED FOR THE DEGREE OF

MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2004

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to everyone who has been involved from the inception to the completion of this research and thesis.

Firstly, I owe my deepest gratitude to the Almighty God who has given me the ability to seek higher knowledge and His divine guidance that has seen me through the difficulties of living and learning in a place away from home.

Secondly, I would like to express my heart-felt thanks to my supervisor Dr. Mohan S. Kankanhalli who has provided me the required direction and support to carry on with my research through its rough spots. His suggestions, comments and valuable guidance have only contributed toward improving my research work and this thesis. I am grateful to Dr. S.H.Srinivasan for his suggestions and for first proposing the idea for this research work. The whole of DIVA group has been very co-operative, helpful and fun to work with. In alphabetical order, acknowledgements are due to Achanta Shri Venkata Radhakrishna, Chitra Lalita Madhawacharyula, Ji Yi, Pradeep K.Atrey, Yan weiqi for their interest in this work and providing helpful discussions along the way. My thanks also goes to other friends and labmates for their company and camaraderie. I would also like to thank the School of Computing, National University of Singapore for providing me with the research facilities to work with during my research.

Finally, I would like to proffer this thesis to my beloved parents and sister. There is no way to express the gratitude to them who though staying miles away, have been a source of untiring moral strength and faith, encouraging me always to give my best and aim higher.

Table of Contents

Abstract

Music in cinema has come a long way since the days of silent movies. Adding sound to movies has revolutionized movies by adding excitement, suspense and right emotional impact producing riveting audio-visual effects, thus lending them the essential fifth dimension. But this creative and artistic ability of professional artists is not available to home video making amateurs. The abundance of home videos and the need to make them more appealing has spurred research in the area of audio-video mixing.

Earlier efforts in audio-video mixing have concentrated on looking at music as reference base to which different video clips are added based on their suitability to produce music digests. The proposed method here is a novel one - of adding music to videos by means of synthesizing music. It is a semi-automatic mixing solution where the users can select music of his/her choice. Then music is suited and adapted to every video is generated. The cinematic heuristics of adding sound to movies gives us the mapping between audio-video elements of home videos. The low level audio elements derive the values from their video counterparts, so that synthesized music is suited to every video. To refine the audio elements into pleasing musical elements further, analogical method is adopted. The musical elements considered for composition are pitch, dynamics and tempo. In this method, examples are provided in order to create musical pitch, dynamics and tempo variations.

Music can be represented as a wave, symbolic notation or a contour. Here, the music is represented in MIDI form. The pitch of example midi music is represented as a contour, which is emulated during composition using sequence based comparison method implemented through dynamic programming technique. The tempo is generated using equal tempered scale and dynamics is varied linearly according to brightness of videos.

The results produced have been tested by users. They are well-received and encouraging, proving that the synthesized music goes toward enhancing video appeal.

List of Figures

List of Tables

CHAPTER 1 INTRODUCTION

1.1 Background

The use and applications in digital video has seen an upward rise over the last few years. The advantage of digital video is that it is easily manipulated and this feature makes it very attractive to consumers as well. As the interest in the use of camcorders and digital camera increases, a lot of amateurs as well as professional people will shoot a lot of home videos. Many digital video editing applications aid in edit video of modest dimensions and integrating them with other media such as audio, photographs or other computer generated images. Though the video can be edited to make a slicker production, video without sound is not very engrossing and appealing. Just as sound and music animated the silent cinema, the addition of audio, video soundtrack or music or both appropriately mixed can result in interesting music videos.

Most of the existing commercial software enables the home user to add music of his/her preference and edits the video according to the music selected. It also assumes that the user has enough knowledge about the aesthetic mixing principles. But this approach may not be successful because the user could be a novice and does not necessarily know about the right principles for aesthetic mixing .Hence the effect of mixing audio with video will not be optimal aesthetically. Automatic audio video mixing is one of the ways to address this problem but without totally excluding the user. Instead of mixing music by feature extraction and subsequent matching , another approach is to synthesize music by 'listening' to the meaning inherent in the video using underlying principles of computational media aesthetics [12] to generate customized music for every video clip.

1.2 Motivation

There are different ways to approaching the problem of mixing audio and video. In the earlier work on audio-video mixing [32], certain features of the video and audio were extracted and based on matching criteria presented in [50]; the best clip for the audio was chosen. This relied on the accuracy of the feature extraction from both media i.e. from video and audio. Extracting features from an audio signal is not an easy and has its limitations. The matching between audio and video therefore left room for improvement. Another way to add audio to video is by selecting a sample audio piece and then using it as a baseline to select relevant video clips based on matching features between audio and video. It derives the basic structure for transitions, cuts and other editing actions from the audio selected [Foote et al]. But we have looked at the problem from a different perspective. The area of media semantics is emerging and therefore this research explores a novel method of accompanying a video with music. The principles of computational media aesthetics gives us certain guidelines for matching audio and video. Based on this information and the examples provided by the user, music is synthesized so that it follows the semantics of the video, thus resulting in music that is suitable and semantically relevant to it.

1.3 Contribution

This research proposes a way of adding audio to video by synthesizing appropriate music based on the video content. The system takes in music examples from the user and generates new music by applying the aesthetic rules of audio-video mapping. To generate music, one needs to consider the elements involved in music composition and generate them. The main elements that are important in music from any culture are pitch, tempo, dynamics and rhythm. From the aforementioned elements, pitch, dynamics (loudness) and tempo variation is explored in this research. Pitch is generated through contour based pitch

matching which is based on ideas from string matching and using dynamic programming as a technique to achieve it. The tempo is varied according to the motion of the video and thus provides variation in the rhythm. The dynamics generation is through variation of the volume or loudness of the music which is controlled by brightness of hue in the video.

1.4 Document Structure

This thesis is organized into five chapters. Chapter 1 is about the introduction where the motivation and contribution are outlined. Chapter 2 elaborates on the required background knowledge and the concepts; rationale behind the approach adopted for research and summarizes related work on music composition in general. Chapter 3 describes in detail the approach used and the algorithm. Section 4 gives the results of the experiments and section 5 gives the conclusions inferred from the results and the future work in this direction.

CHAPTER 2 LITERATURE

2.1 Audio Video Mixing

Audio video mixing is an important part in movie productions and sitcoms as this gives the required emotional and aesthetic appeal to the video. Software such as Muvee[36] gives the user the opportunity to mix the audio clip of his choice with the video and accordingly aligns the video clips to the chosen music introducing some special effects such as gradual transitions. But the effects of this would not be optimal if the user is a novice and has limited understanding of the aesthetics rules required for professional mixing. In [32], the authors have presented a 'pivot vector space method' of automatically mixing the audio clips with the video suitably taking into consideration the aesthetics aspects. The video aesthetic features are extracted from the low level video features such as color, light falloff, motion vectors. The audio perceptual features such as dynamics and tempo features are calculated from the low level features such as spectral centroid, zero crossing rate and the volume. Matching the video with the audio segments in the pivot space lies in computing the best match for the each video shot and music excerpts from the Euclidean distance and then the overall best match. Another interesting application of audio-video mixing is the automatic creation of music videos as presented in [20] by Foote et al. Here, the home videos are aligned with the selected musical work by automatically segmenting the audio, video, finding an audio novelty score which indicate the peaks in the audio track and then aligning the peaks in the audio with the video clip boundaries. This creates a digest of the home video by removing certain clips that are found unsuitable with regard to camera motion and video brightness. As mentioned in the paper, synchronization between video and audio segments has the effect of enhancing the visual perception. This paper considers the audio-video matching from the point of taking the audio as a baseline on which a video

montage is built. But the suitability of the music itself for the particular video is not taken into consideration, as the rhythm and style of the video digest is driven by the pace and rhythm of the music. So a slow paced music gives a style different from that of a fast paced one. But one can also view audio-video mixing from the point of the video being a controlling element, instead of the audio and appropriately choose and control the audio elements to match the video. This is the underlying idea used in the research here.

2.2 Media Aesthetics

Media aesthetics is the study of visual and aural elements, the interaction and integration of these elements to understand the semantic and the semiotic content of the video by Dorai et al [12]. It explains the basic aesthetic elements of media that fit into contextual fields, analyze their interdependence, thus clarifying and suggesting how to match pictures, sound effectively. The five fundamental elements of T.V, film production are light and color, 2-D area, 3-D depth and volume, 4-D time-motion, 5-D sound. Studying the aesthetic characteristics would help us answer the questions on aesthetic matching between image and sound. Elaborating on these elements: The interplay between light and shadow communicates with our inner emotions example the swaying of long shadows of the trees in the silent night creates a picture of eeriness in our minds. The direction of light is also important. If the object is illuminated from below, the attached shadows fall upward as is often used in horror scene lighting. The five dimensions used in aesthetics of image and sound are described below: [12]

Light fall off: A fast light fall-off emphasizes volume and texture and is used to heighten dramatic scenes where as a non-directional, soft light creates the opposite effect.

Color: Color is used to establish a mood or add excitement to an event. Most film directors set-off a high energy foreground against a low-energy background. High energy

foreground or background is produced by using high saturation color like red, low energy with low-saturation colors such as light green, azure blue. Psychological, introspective movies such as Schindler's List rely on de-saturated colors or the total absence of colors to match the mood and context in the scenes.

4-D field(Time Motion): Time in a movie has to be controlled to make a viewer perceive a desired pace and rhythm. For instance , in a car- chase scene, the director may rely on fast, short cuts to give that effect instead of actual speed of the cars.

5-D field(Sound): Sound is divided into literal and non-literal sounds. Literal sounds refer to the sound -producing source on screen where as non-literal sounds are not connected with any object on screen. Music is a non-literal sound that intensifies the energy of the event on screen. Like color, music directly affects our emotions. Music can be used to impart a rhythmic effect to the shot sequences. If the shots are unstable and shot sequences aligned erratically, then a rhythmic sound induces a rhythmic perception in the audition.

Manipulating sound energy in the sound track also adds a high level of impact and energy to the video content. Sound has outer orientation functions that relate to the spatial temporal aspect of the video as well as inner orientation value that signify the mood, energy and structure of the events. Sound energy is associated with different meanings. In horror movies, the sudden shift in sound energy from low to high gives a dramatic effect of shocking a person and the return back to low energy evokes a release of the tension built. Sound can be added in different ways. If the rhythmic structure of the sound moves in parallel with that of video, it results in a stable and tension-free perception of the video. The effect of counterpoint, a word derived from music (means playing simultaneously one melodic line with another), in a movie can is used to highlight certain scenes in the movie. The extreme violence and destruction of war scenes, when contrasted with absence of music, creates a lingering, contrapuntal effect in the audience.

Tempo: This refers to the rate of delivery of information in the film. It tells us whether the movie moves swiftly or tardily. It also affects the perception of time of the audience. The tempo of film is increased by increasing the object motion, camera motion, 'motion' imparted by film editing. In [12], Dorai et al have considered motion, shot rate and sound (dialog, score) for tempo calculation. Motion is considered to bring in new information into the consecutive shots or it may just rack objects without bringing in much of information. The other important factor that affects tempo is the shot length. According to Zettl, [50], a shot length manipulation has an effect on the perception of event density and thus affects subjective time (tempo). Even if the shots are rapidly cut with very little new information, it captures the viewer's attention till he adjusts to this rate of delivery. If the shot lengths are short the tempo is staccato in nature but if they are long and motion slow, then it is a legato style. This similar to tempo descriptions of music. Rhythm is also derived from the inherent structure in the video. Video rhythm is divided into shot and motion rhythm. Motion rhythm is again divided into metrical, attack, decay and free. The tempo and rhythm described above have corresponding parallels in music. Musical rhythm is divided into groups and groups into measures. Meter is split into bars and every note is determined and defined by the onset time (attack), the duration of the note and the decay time.

Tempo function: The work done by Dorai [13], fixes a lower limit to the length of shots. The tempo weighting function is given which indicates that the most of the shot lengths are clustered around the median of all shot lengths in the video and the number of shots rapidly decreases as the shot length moves way from the median value. The median is the zero point of deciding the tempo of the shots. It is found that tempo is sensitive to the change of shot length around the median. Based on this tempo weighting function W(s), the tempo equation is calculated.

$T(n) = \alpha(W(s(n))) + \frac{\beta(m(n)-\mu_m)}{\sigma_m}$ [13].

S - shot length in frames, m - shot motion magnitude n - shot number. $\alpha, \beta = 1$ since both the shot length and motion are weighted equally. W(s(n))- weighting function for shot length normalization. The motion magnitude is estimated using the algorithm in [30] where the medium to large object motion is captured along with the camera pan and tilt in two consecutive frames. The motion estimates which are away from the mean by 3.5 standard deviations are not considered and the result is smoothed by a Savitsky-Golay filter to dampen noise[46].

Generally, the change in plot of story or change in the location or onset of a new phase in the movie results in the change in tempo but this may not be necessarily true.

Herbert Zettl in [50] has highlighted interesting ways in which video can be combined with audio. To make a audio-visual combination, the video vectors and audio vector fields should be combined synergistically. The way to combine synergistically is to create picture-sound unit which results in an effective gestalt. Zettl gives guidelines to make the structural matching as given below

1. Homophonic structures: Each audio event runs parallel to the video event. Many music videos are created with sound track accomplished by relevant video shots. As the scene changes, the audio track also undergoes a change.

2. Polyphonic structures: Picture and sound have independent 'melodic' lines yet combine to produce an intensified scene. Phasing is a common way to create a polyphonic structure i.e. audio event of one shot extends into the next or more that one video event is accompanied by the same audio event.

3. Montage: Interesting audio - visual combinations can be created by combining music that is either in adds to the homophonic structure of the visual elements or by selecting music that is in contrasts to it. For example, playing a soft and slow music

against the backdrop of violent scenes in the video creates a collision montage.

4. Picture - Sound matching criteria: It is difficult to match video with music or vocals that are appropriate to it to create the right aesthetic effect. Some of the different ways in which to match video and audio are:

   (a) Historical - Geographical matching: To match video with music that was prevalent at the same time, such as the dances of the medieval period in a period film is matched with music of that age. Geographic matching takes into consideration the region and the regional music for matching.

   (b) Thematic: The matching is done based on the theme expressed in the video, such as the video of a football game is matched with the music of a lively band.

   (c) Tonal: The matching is done based on the mood or feeling of the event on screen.

   (d) Structural matching. The match audio and video based on the internal structure of their elements. In order to find the most appropriate music to the visual event, we need to identify the video vector fields and the audio vector fields and then use the contextual variables to give the desired effect. A table of such a structural mapping is given in Zettl [50]. Structural mapping help to answer questions such as, is the image of a rose gentling swaying in the breeze simple or complex in structure, does it 'sound' soft or harsh?, etc.

2.3 Interplay between Audio Visual elements

The addition of sound on image can create varied effects. The sound can directly participate in the scene by assuming scene's rhythm, tone and phrasing. This is called empathetic sound, which is derived from the word empathy [11] meaning to feel the feelings of others.

Sound can also be employed to create a visual impact by exhibiting complete indifference yet producing a riveting effect in the viewers mind. This is called an anemphathetic sound. Citing an example from [11], in the movie Psycho, the violent murder of a character is followed by the 'ordinary' sound of the running shower. There are other kinds of music that are neither empathetic nor anemphathetic but merely serve as a 'functional' filler with makes no emotional meaning in it.

It is interesting to observe how sound is employed to create a perception of time on the image. One of the aspects that sound brings to image is temporalization. Sound can be used to temporalize an image in 3 ways.

1. Temporal animation of image: If the image consists of static shots or an image with only fluctuation in movement, then sound infuses a sense of temporality into it. For instance, the image of rippling water can be enlivened by a sound of dripping water.

2. Temporal linearization of image: If the image has temporal animation in it already, then sound brings about temporal linearization in it. The sound combines with the visual elements to produce a concerting effect or results in an opposite effect. For example, the footsteps of a person climbing the stairs when accompanied by the sound of it produces a unifying effect in our mind.

3. Tempo: The acceleration or deceleration of the pace in the sound does not necessarily impose a similar perception on image. Music played rapidly does not actually increase the speed of viewer's perception of the image. Instead if the temporal animation will be greater if the flow of notes is unstable yet moderate in speed. A regular pulsation of sound has less temporal animation than if the sound is irregular. If a constant stream of notes is played on the violin, it becomes difficult to fit the images into such music. But if there is music that is played with a regular rhythm that is

periodically repeating then it creates 'expectation' in the viewer as he awaits the onset of particular events in rhythm. This influences the pace at which the moving images are observed. It has also be noted that sound with high frequencies helps to capture the viewer's attention.

Another aspect of sound that animates image and is of importance in movie production is the synchronization of the sound and image. This is brought about by distribution of synchronization(sync) points in the flow of image sequences. This can be observed clearly when the sound assumes the same vibrating, trembling quality of image. Such dramatic movements are called microrhythms on image. For instance, a candle flickering in the wind, ripples in the pond, petals showered from a tree also evoke mental pictures which have a rhythmic texture to it. If the sound added to these images agrees with the fluctuating quality of images then it forms a good sync point. This definitely serves to enhance the visual appeal of the image.

2.3.1 Points of synchronization

As mentioned earlier, sound has the effect of intensifying a particular action or movement in the image by following the same rhythm as suggested by the visual element. A sync point in the audiovisual sequence is a salient moment during which a sound event and visual event meet in synchrony. This leads to something referred to as synchresis in Chion [11] (word formed by fusing synchronization and synthesis). Synchresis is the spontaneous fusion which is formed when certain visual elements in the image meet their audio counterparts. Such sync points generally obey the laws of Gestalt psychology which shall be discussed in a later section. The following illustrate the meaning of synchresis:

1. A sync point coinciding with the close-up of a shot thus creating the effect of visual fortissimo which can be again heightened by a gradual increase of volume along the

length of close-up.

2. A synchronous cut between sound and image track such as a visual cut at the end of a spoken sentence.

3. A dialogue spoken emphatically can act a hinge for sync point with the character in the image.

Meaning and rhythm can bring about sync points of agglomeration. The phenomenon of synchresis is used in many ways by film producers for dubbing and sound effects mixing. In case of dialogue synching, one finds many sync points but only a few are important. This leads to audio-visual phrasing of the dialogue and image.

Until now, the focus was on how sound can transform and animate image, but the reverse kind of phenomenon can also exist i.e. image can be used to 'magnetize' sound. Earlier in monaural films the sound issued from the only one loudspeaker that was stationary. The point of sound emission did not move as the source moved on the image. So if a person ran from one side of the screen to the other, the sound of running footsteps didn't follow the source. The psychological localization of the viewer is relied upon to connect the actual position of the source and a different localization of sound. Picture the scene of a person far away in the distance located on the left side of the screen calling out over the grasslands. The distance of the sound source can be interpreted by the viewer through mental spatialization and thus localize the sound in space. The sound source does not have to be placed very far from the screen to produce this effect. This mental localization occurs more by what the eye sees than by what we hear and also due to the combined effect of both.

Sound can be divided into three zones the way they are applied on image as illustrated in the figure below[11].

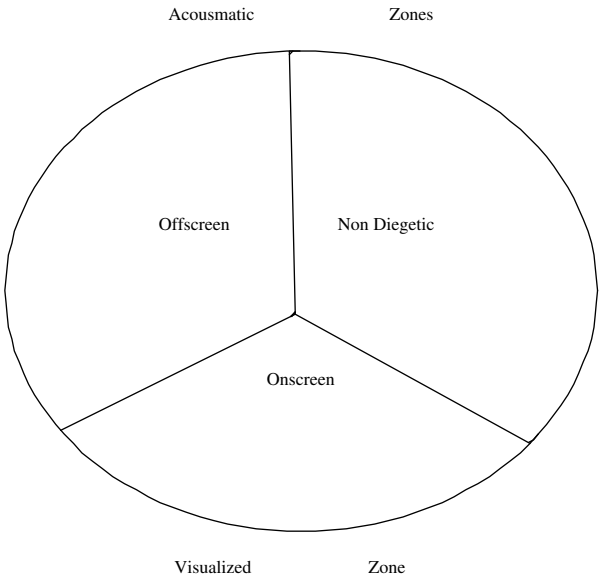

Figure 1: Three zones of sound

Offscreen Sound: The object producing sound is not visible on image. E.g. We hear the birds chirping but see the picture of a child waking up in morning.

Onscreen Sound: Source of sound appears on image. We hear the dialogue between two people and we see them on screen.

Non-diegetic: The source of sound as well as sound is external to the image and also to the story such as music underscoring.

Music plays a special role in the audio visual contract. It is not constrained by the need to be related to the diegetic space and time as other sound and visual elements are. It makes the visual elements more malleable as it stretches or contracts the events on the image. For example, the capacity of holding a suspense moment on screen by freezing action yet continues the music in the background.

We find interesting applications of engaging sound and image in animation films. To understand this, consider the example of a cartoon character that climbs steps. As he

climbs, the music follows the pattern of rise and fall of footsteps yet the pitch doesn't rise in scale levels. The trajectory of the sound is imitated. This pairing of music and image is called micky-mousing where the visual action moves in synchrony with musical trajectories (rising, falling, zigzagging and the punctuation of sound). As we draw examples from everyday life and apply this to the fusion of sound and image, we observe that image can be re-associated with different sounds from different sources yet not detracting from the meaning or essence of the image. For instance, the sound of axe chopping the wood can be applied with synchrony with the exact moment when the cricket bat hits the ball thus creating a sync point.

2.3.2 Gestalt laws and Music Perception

Gestalt psychology is the psychology of perception which is also called as perceptual organization. Gestalt laws of psychology were first applied to visual research and then extended to music. We all possess the ability to fill in the gaps in visual information to complete patterns and configurations. This perceptual activity is called psychological closure. For example, when we see, a set of three dots arranged in the shape of a triangle, we perceive a triangle rather than three individual dots. The pattern that results in closure is called as gestalt which is a perceptual whole that transcends the parts. In music, we can take the analogy of chords which is composed of notes sounded together. This is a gestalt as the perception of a chord is very different from that of three notes sounded in sequence as in a melody.

The gestalt principles of proximity, similarity and good continuation have been used in the model of Narmour (1990) [33]of implication-realization that aims to answer questions on melodic expectancies. The underlying principles in this model is based on the gestalt principles. The implication realization model describes music cognition as a succession of

points of closure and points of implication. The expectancies of melodic continuation are weak at points of closure of rhythmic, meter and harmonic stability but at the points of implication where there is instability in rhythm, meter and harmony the expectancies are strong. The principles of gestalt have also been applied to cognitive grouping of musical events when grouping events according to rhythm. The proximity principle, objects which are close together and the similarity principle which is applied to objects that are similar are perceived as groups. These principles of similarity and proximity are responsible to bring about a sense of cohesion in music in the listener's mind.

2.4 Basics of Music Theory

Pitch: The pitch of a sound is determined by frequency. This indicates the relative highness or lowness of sound. A sound with a specific pitch is called a tone. The distance in pitch between two tones is called an interval. Tones that are separated by an octave sound similar. Pitch can be used to create musical moods Kamien [24]. Low pitches signifies solemnness but high pitch signifies brighter moods. In case of western classical music, definite pitches are used but African music played on drums consists of indefinite pitches played in a rhythmic fashion.

Dynamics: This indicates loudness of music. Loudness is related to the amplitude of vibration of the sound. A steady increase of dynamics stirs up excitement that is usually accompanied by increase in pitch. The gradual decrease in dynamics suggests calmness of mood. The dynamic level of a tone cannot be indicated absolutely. To indicate gradual change in dynamics, the following words are used: Decrescendo or diminuendo for gradually softer and crescendo for gradually louder.

Tone Color: It refers to the timbre of the music instruments in the musical piece. The same melody played by different instruments sounds different because there is a change

in tone color. A change in tone color is generally used to highlight certain movements in the music. It is also associated with certain contexts and meaning. Trumpets are used for heroic tunes or a flute to produce soothing music.

Key: Key is important in melodies since most of the melodies are composed around a central tone. Key also involves a scale and chord built round the central tone. The seven tones in music are given by C D E F G A B which gives the basic scale. If the music is played in the key of C, then the opening note is C and that becomes the tonic, the basic chord being the triad C E G.

Scale: A scale consists of basic pitches of music pieces arranged from low to high or high to low. In western music, the basic scales are a) Major b) Minor.

Major scale: It consists of 7 notes that are sung at seven different pitches and arriving at the first note which is an octave higher than the first. Octave is the distance between first and eighth tones of major scale. The intervals in this scale are half steps and whole steps.

	WS		WS		HS		WS		WS		WS		HS
C		D		E		F		G		A		B	C

WS whole step, HS half step.

The scale shown above is the major scale with C as the tonic .As the distance between B and C in the end of the scale is only a half step, there is a strong tendency to pull towards the tonic (the central tone) at the end. There are twelve possible major scales depending on the twelve tones in the octave. The same pattern of intervals is followed for any beginning note of the scale.

Minor scale: It consists of 7 different notes with the eight duplicating the first an octave higher which is similar to the major scale. But the pattern of intervals is different which

produces different music.

	WS	HS	WS	WS	HS	WS	WS
C	D	E	F	G	Ab	Bb	C

WS -whole step, HS half step.

Music based on minor scales is more serious and sound 'darker' and music in major scales sound brighter. Various combinations of scales may be used in music composition to provide more variety and create mood contrasts.

Rhythm: It refers to how the music ebbs and flows against the passage of time. It is expressed by beats, meter, accent and tempo of music. The note lengths are usually varied by setting them against the timeline of beats.

Tempo: It refers to the speed at which the beats are played. It is specified in beats per minute(BPM). There are different scales on which musical tempi is based such as harmonic scales, equal tempered scales, metronome scale, augmentation/dimunition scales. We have adopted the equal tempered scale in our research.

Melody: Melody is determined by combination of pitch series and rhythm resulting in clearly defined shape. The duration of notes and their ordered succession of intervals define a melody. Melody may start on note C, rise up to a note an octave higher, then come down to the starting pitch thus following a melodic arch or contour.

Harmony: It is composed of chords and is based on the progression of chords. Chords consist of tones that are sounded simultaneously.

2.5 Representation of Music

Music has symbolic and structural relationships between the different dimensions such as pitch, time, rhythm, tempo, timbre etc. It can be treated mathematically for analyzing

elements such as pitch, rhythm etc but there are also non-mathematical aspects such as emotion, expectancy etc as stated by Dannenberg [39]. The musical representations depends on what problem we are trying to solve. The non-mathematical aspects can be best studied by perceptual modelling of music which looks at music from the cognitive point of view. The cognitive representations aim to understand music as humans perceive it. The basic representations are based on symbolic and structural information.

In the overall representation of music, music is considered to be composed of a hierarchy of levels from the highly symbolic denoted by printed music to the non-symbolic represented by audio signal. Each level contains information that is not present in the other. Music symbolic notational representation includes the representation of music structures such as key and time signatures, slurs, rests etc. and also graphical information such as staff position etc. This notation has problems in computer music applications as this information is mainly visual and all the score information this cannot be effectively processed. In recent years there have been graphical editors that process the musical parameters and other solutions to the issues in music notational systems but there are still many open issues[39].

Musical information is divided into continuous and discrete data. Continuous representation incorporates changes over time and is usually represented as splines (piecewise linear functions) or by mathematical functions. Discrete data represents events at a point in time. In music research based on neural networks, the pitch is represented in terms of frequency or a set of triads containing the pitch and this information is distributed over layers of the artificial neural networks instead of being stored in discrete data structures.But according to [39], the discrete data is advantageous to music representation as there is a natural correspondence to musical notes which have start-time and intervals and hence easier to process.

2.5.1 MIDI representation

The musical surface is a collection of notes or playing instructions that is presented to the musician through a written musical score or a MIDI presentation. Midi is often used as the prevalent standard in music composition systems and is also used a transmission protocol. It is an acronym for Musical instrument digital interface, Midi originated as a hardware protocol that enabled communication between electronic keyboard and computer sound applications. But MIDI software packages and MIDI IO libraries use the standard MIDI file format specified in ASCII format as the input. The standard Midi file specification can be specified in three parts: Header chunk types, track chunks and event types. The event types are divided into midi events and meta events. Midi events are divided into note on/note off events, time clock, timing standard etc. Note events contain five items of information 1) Delta time 2) Note on/note off status 3) Channel number 4) Pitch (note number) 5) Attack velocity (dynamics). Meta events contain track names, tempo indications etc. Delta time indicates delay time between onsets of discrete Midi events in a serial stream of data. Delta time indicates the amount of time in ticks that has elapsed since start of last event. The delta time format and the initial tempo are specified in file's header. Since Midi protocol is a message oriented one, when sound begins, a note-on message is sent and when the sound ends the note-off message is sent. The channel number is used to specify the instrument to be played in that channel.

Pitch is represented by key number. The MIDI note scale is in the range of 0-127. Middle C is assigned to key number (e.g. 60). All the numbers that are multiples of 12 are Cs. There are 128 notes in MIDI notes out of which 88 notes are the piano notes, the remaining are 20 additional bass tones and 20 additional treble frequencies. In MIDI output, tones are present in octave corresponding to chromatic scale. The black notes of piano specified as sharps C#, D#, F#, G#, A#. Key velocity or the dynamic range is given

on the logarithmic scale. 1 - ppp (pianissimo), 64- mp(mezzopiano), 127-fff(fortissimo). 0 is reserved for signal Note-off condition. Tempo is specified as a header and so is the time signature which tells how many quarter notes, eight notes, sixteenth notes, thirty second notes have to be played. The defaults are for tempo and meter are 120BPM and 4/4 meter.

Header chunk contains information such as ticks per quarter notes, number of tracks in the song etc. Header information contains key signature, time signature and ticks per quarter note. [See Appendix B for MIDI file format in ASCII.]

2.5.2 Melodic Information Processing

This considers representation of melodies from a psychological view point. Musicians and psychologists have found that melodies can be repeated at different pitch levels. But the identity is still preserved as long as the pitch intervals between notes have been preserved through transposition to a new key along a log scale of frequency. It has also been found that preserving melodic contour (pattern of ups and downs) while allowing interval sizes to change (through tonal imitation) produces something that is similar to the original but not identical to it. If the melodic contour is maintained while changing the interval sizes more radically, then the result is atonal. If we alter the melodic contour along with interval sizes, while maintaining the rhythm then the music that results is representative of the broader class of similar songs [16]. The various different patterns of music described above evoke different responses in an adult's mind. The mental abilities of an adult to recognize music similarities and differences in music differ from that of a child's brain.

Dowling [16]has explained the evolution of melodic processing from a child to and adult from the perspective of memory retention of the features of melody. What features are selected, perceived, stored depend on the age and culture of the person. At the local level, the features perceived are pitches and durations of individual notes while at the global

level the features are whole phrases, contours, rhythmic pattern and the tonal scale. The choice of features perceived also depend upon the culture. In western music, the contour can be changed independent of the mode of scale (major or minor) but this may not apply to other musical cultures. Choosing a particular contour restricts the choice of each succeeding pitch. When a particular pitch of the note is selected from the range of pitches, the next note to be taken depends on the direction of particular interval in contour. When subjects were asked to distinguish between exact transpositions of melody to a new key and imitation at a pitch level but having the same contour, they were unable to distinguish correctly between the two, unless they had listened to the melody many times. Contour as well as interval information is difficult to remember over a long term. But despite this fact, the contour information is easily encoded in memory but interval information is difficult to encode. Repeatedly listening to the melody can help retain the interval size information. Contour information can be used as an " indexical device to access melodies in long-term storage, but recognition of such melodies seems critically dependent on scale step information " [16]. To highlight the importance of contour, the following example is shown [16].

The melody 'Twinkle, Twinkle little star', can be encoded in terms of sequence of ups-downs signs for contour representation or else by the number of diatonic scale steps.

In Semitones: 0, +7, 0, +2, 0,-2. In diatonic scale steps: 1, 1, 5, 5, 6, 6, 5 in terms of degrees of major scale.

The listener is able to match the contour (sequence of signs) quantities but does not match the quantities (interval sizes) when a tonal imitation is done. But when fewer pitches are shared between the original and the imitated contour, then interval sizes regarding which some information is stored in the long term memory, can be used to recognize the original.

All this proves that, whether the imitation is tonal or atonal following the contour of the original, maintains similarity with the original melody and unless the interval sizes are radically changed, the new piece of music generated based on imitation is connected in the listener's mind with the original. The interval sizes are recognized with comparative ease by a trained singer rather than an untrained one although untrained listeners' have a notion of it and training helps them to apply it in recognizing tonal structure (Krumhansl and Shepard 1979). In the research done here for pitch sequence generation, melodic contour is used for the generation of pitch in the experiments explained in section §4.4.4.

2.5.3 Contour based Music Representation

The word melodic contour in the simplest sense means the "up-ness or down-ness of the notes in melody". It conveys a sense of continuous, flowing motion in music. It is described as a mid-level representation of melody between sample data and melody model. It is neither a low-level representation like a sound wave nor a high level representation like the discrete and symbolic notation in music. Most of the low level systems in music recognition are data driven which analyze the pattern from huge data sets and not much high level knowledge is used.

Dowling [14] has described two components which contribute toward reproducing and recognizing music. They are musical scale and melodic contour. He points out that people remember melodies as a sequence of pitch intervals between successive notes. He also gives evidence to the fact that the melodic contour is remembered separately from absolute pitches or exact interval sizes. When a listener is trying to retrieve the set of intervals of melodic tune from his memory, he relies on the contour to retrieve the melody.

In music cognition, contour is represented by a sign that indicates whether a note if higher, lower or same as the previous note. The ternary representation (+/-/0) has been

accepted as a standard definition of melodic contour [6]. According to Dowling, inexperienced listeners represent melodies as a sequence of intervals and more experienced listeners as a scale-step representation. The ternary representation though good for coarse representation, is not very effective in capturing the richness of melody as it doesn't indicate the interval size or take grouping of notes into account. Two sequences which are different may look the same in ternary representation as the exact interval sizes are not specified. But despite the inadequacy of ternary representation, experiments conducted on melodic recognition on melodic contours using ternary representation show that pitch interval direction (represents rise, fall, remains constant)is important for melodic recognition. And a coarser melodic contour description helps listeners to determine melodic similarity. As listeners become more familiar with melody, intervallic distances take greater perceptual significance.

A melody can be uniquely identified even after it has undergone transposition (i.e. we can still recognize a tune that is sung in a different key or a scale higher)[18]. Absolute pitch doesn't represent melody accurately. It is more useful to represent difference in pitches as intervals between one note and the next as intervallic relations are invariant to key transposition. Contour is a subset of intervallic information and so is also invariant to transposition. The musical element that is not invariant to transposition is rhythm which corresponds to occurrence of notes as specific times in relation to the meter.

An alternate way to represent music is through structural approach which is not described here. A hierarchical structure of music is described in Lerdahl and Jakendoff[1984].

2.6 AI and music composition

Artificial techniques have been often used to learn the expressive interpretation of music pieces which involves learning musical parameters such as the dynamics (variations in loudness) and rubato (variations of local tempo) (Widmer[47]). Many computational intelligence techniques have been used to solve musical problems like music cognition, algorithmic composition, and sound synthesis (Burton and Vladimirova[5]). An overview of some of these techniques is given below.

We can look at the art of music composition from the two points of view. One is to make computers compose music in order to help composers and another is to make music composers. The former approach had led to many experiments in the combined field of AI and music cognition while the latter is still a very open area. One of the first to experiment with music composition computationally is Xenakis [48]. He tried to produce scores of live ensembles through probabilistic and statistical methods. His program would generate music given a list of notes densities and probabilistic weights supplied by the programmer.

People follow different view points for composition of music computationally. We can look at composition from the perspective of a music composer who writes the score or from the perspective of a listener. Music composed as a human understands it is based on the music cognition. The task for a music program based on music cognition is to make the output of the algorithm to match the performance of the human composer by emulating the human performance. But the music produced though reasonably good may not necessarily be something the way a human understands it.

Artificial intelligence techniques have come to aid of music composition by opening up possibilities that make use of understanding of music cognition as well as rules of music theory for composition and thus can support a wide variety of models for music

cognition. The research in music based on AI techniques falls into two categories. 1) Rule -Based systems 2) Model -Based systems. This is again dependent on how we view music composition. Do we consider music as a form that is determined by rules set down by the ancient composers or do we view it from the listener's perspective where the ear guides what is perceived as musical.

The connectionist approach to music follows the second principle where a bedrock of rules does not determine the musical structure of the composition. Instead a model is built out of examples that train it and the training in effect lets the model deduce its own rules. It is not constrained by the rule base as in rule-based systems and thus offers more flexibility in producing novel compositions within a limited solution space. New material is produced by the model after training by exploring the whole solution space constructed by the examples. Usually connectionist methods employ neural networks which after adequate training can be analyzed and regularities in the examples can be deduced. This post training analysis obviates the need to create a formal specification for rules. Barucha and Todd [28], trained neural network to understand and anticipate common musical practices but also showed that the network is capable of generating exceptions to the norm of formalized music.

Connectionist models for tonal analysis exist where the network algorithm trains the network based on note durations, order of the notes along with the number of times that note appears in music [28]. The model consists of separate units for key nodes, chords and pitch class nodes. The notes in musical sample activate the chord nodes which then transmit the activation to the key nodes which are chosen depending on whether the chords are tonic, dominant and sub-dominant. The representation of pitch is also made possible through the use of neural network.

Rule Based system: Hiller et al [21] thought of music composition as a problem of prescriptive rule-based composition. They used perspective rules to discover valid music

sequences from the space of all the musical sequences. But difficulty lies in getting the whole set of rules of all musical forms in quite a huge task. Though rules form the foundation of composition, all the various musical styles of different cultures cannot be encapsulated by a set of rules. For a listener to appreciate music, expectation in very important. An artist plays at different levels of a musical structure, combines rules and creates new ones sometimes violating the standard set of rules in order to hold the listener's attention. So, generating novel arrangements are a challenge to rule -based systems.

Expert systems for harmonizing chorale in the Bach style by Ebcioglü[28] or expert systems for harmonic analysis of tonal music by Maxwell[28] have been based on use of constraints and heuristics. These constraints and heuristics are essentially derived from treatises of harmony and counterpoint. But great composers do not strictly follow the rules of the book and break these rules to introduce variation. Usually compositions are written incrementally from left to right and at each stage of composition composer chooses an item (e.g. a chord, a phrase etc.) to add to the partial composition. Along with the enunciation of absolute rules which describe real music, one must also find heuristics that tell us about possible ways of extending the music piece to produce musically pleasing pieces. If the heuristics are not used, the solution space populated with unmusical patterns and then application of complex constraint becomes necessary.

Genetic algorithms: Genetic algorithms (GA) have been mainly used to solve compositional and synthesis tasks. These algorithmic composers operate on music knowledge such as pitch, rhythm, meter and on the rule representation which contains a set of rules that determine how the composition evolves [5]. GA are especially useful when generating improvisations and producing variations of already existing music. The composition by GA depends of three areas. First is to build a search domain that will give rise to a multitude of combinatorial possibilities in rhythm, pitch, individual notes and note durations. But

we need to prune this vast space by imposing constraints which limits the generation of music restricted to a range of notes or a particular key. Secondly we need to represent the musical knowledge in the form of pitch, rhythm and meter. Thirdly the fitness functions that decide how the composition evolves over time. So the fitness functions form the crux of the success of GA as this forms the basis for choosing the individuals from the population and using them in the composition. GA is used in harmonization of melodies or used to produce jazz solos over a given chord progression. [9] has come up with a system called GenJam to create Jazz solos. This uses interactive GA (IGA) to evaluate the fitness function for each of the phrases and measures which make up these phrases by a mentor. But the success of GA in composition largely depends on the fitness evaluation function which is subjective in the case of user defined one or is based on rules of music theory. Since the search space is unlimited, it becomes a difficult job for a user to evaluate all the measures and phrases in the music samples and the other disadvantage with this method is that rules need to be properly represented as constraints and this can lead to potential loss of some interesting solutions. It has also been found that genetic algorithms can tell us very little about the mental processes involved in the composition of music thus limiting its capacity to improvise greatly on the music samples.

Machine learning method: In [17], the authors have attempted to generate music by style modelling through statistical and information-theoretic tool, which imitates the style of masters'. The model utilizes two theories, one is the incremental parsing(IP) and prediction suffix trees (PST). Predictive theories are based on understanding musical expectations of recent past context that guides musical perception. Earlier Markov models were used for music generation but the music generated is not very pleasing as either it does not resemble the music it is trained or it replicates the input.

IP and PST methods are dictionary based methods that build a model based on the

phrases or patterns of the input patterns. They provide inference rules to predict next musical objects based on past context. A dictionary of motif sequences is maintained as the sequence is parsed from left to right and new sequences are added to dictionary. PST is similar to IP but builds dictionary of motifs based on the occurrence of significant motifs alone.

Expressive variation of Music: The term expressive variation refers to the shaping of music by varying musical parameters such as loudness, tempo by speeding up or slowing down, growing softer or louder. The music is perceived by listeners through different structures and at various levels. Each musical structure is associated with its own expressive shape and the performer varying this shape emphasizes and de-emphasizes some structures and plays the music as he understands it. Listeners do not hear discrete notes in a performance but organize the events into groups such as phrases, motives etc. As given in [7], the expression decisions are not function of single notes but depend on large-scale structures such as increasing or decreasing the tempo of a musical phrase towards the end. In [47], a given musical structure can be converted into an expressive shape of dynamics and tempo. The rough trends in the curve are identified as prototypical shapes such as ascending, descending, ascending-descending etc.

Expressive variation deserves a mention in this research as the composition of music by analogies is done by taking an example music piece and playing it after changing the musical parameters such as pitch, tempo and loudness. This is analogous to the playing of a musical work by a performer by expressively varying the musical elements. The composer would like to explore and unravel the various possibilities of music creation without necessarily explaining them or codifying them as rules. In [39], Dannenberg has developed an algorithm using dynamic programming which aims to solve the problem of score following in real performance by finding the best match between the actual solo

performance and the score stored in the computer. This best match produced can be used to provide the accompaniment which is already composed and stored in memory to the solo performer. This gives the solo performer the flexibility to change the speed of the score and the best (longest) match is produced that triggers the right event from the stored list of accompaniment events. The algorithm used for pitch matching in our system has borrowed ideas from this process and from the method of string matching which will be explained in a later section.

2.7 Analysis and conclusion of the literature review

2.7.1 Audio-Video mixing

1. Pivot Vector Space Approach (Mulhem et al [32])

   It is based on matching the low level features of audio and video. The difficulty in extracting the audio features limits the mixing quality. The result may not be very impressive for all videos. It is also limited by the number of audio content in the database.

2. Computational Media Aesthetics(Chitra Dorai et al[12])

   Presents the idea of computational media aesthetics that analyzes the video low level features to understand the higher level semantics and semiotics of the film.

3. Audio -Visual Contract(Michel Chion [11])

   He describes the interplay between audio and video elements on the screen, the effect one has in enhancing perception of the other. The interesting point mentioned here is the point of synchronization which exists in many videos such as the use of audio to intensify a particular action or fading out of music as the shot fades away. This increases the coherence between the two media.

4. Applied Media aesthetics(Herbert Zettl[50])

   He provides the principles used in media productions for matching audio elements with video. Please refer to table for structural matching in Appendix.

5. Music Video Creation(Jonathan Foote et al[20])

   Music videos are created by finding the novelty score of audio based on the regions of similar energy, unsuitability score of video based on camera motion and aligning the audio segments to video clip boundaries depending on the unsuitability score.The criteria for matching has been mainly the motion in video and the energy of musical passages. Other aesthetic elements in music and video are not considered.

The study of elements in media aesthetics leads to a holistic understanding of combination of audio-visual elements for producing aesthetic video/film productions. In this research, we have worked with video elements such as color, brightness, motion and corresponding audio elements which are pitch, loudness and tempo. In [12], the tempo parameter has been derived from motion and is directly related to it. But in the work here the note duration parameter has been derived such that it is inversely related to it [explained in §4.2.1].

The study of inter-relation between audio and video elements helps us understand how sound is almost indispensable to evoking an appropriate emotional response to the video/film and creating a lasting impression on the viewers' mind. As mentioned in the review earlier, this research makes use of the non-diegetic sound i.e music, to increase the visual appeal of the home videos. Synchronization or else called as synchresis is an important theme in film production. The research here involves the generation of music where tempo which constitutes one of the elements is altered, following the activity in the video. The music generated has a kind of homophonic structure which derives its form and

structure from video and also the musical surface in terms of pitch follows the video closely. The concept of temporal linearization influences the joint perception of audio and video and this combined with tempo triggers expectations based on this perceptual understanding.

2.7.2 Music Perception and Representation

Gestalt laws of psychology have been applied to understand music perception and cognition. Music analysis or composition involves more than the understanding of the basic elements of music as separate, disconnected events. It also involves the perception of it by listeners, experienced or experienced and also by the musicians themselves, since they do not perceive music as a string of unrelated events but rather as a cohesive group of notes, phrases etc. For example, the ascending or descending line of a melody contour is heard as one group. So, to get a deeper and complete understanding of music composition, one needs to dive into cognitive musicology and music theory.

Music can be represented in three ways: symbolic notation, the signal level (waveform) representation and mid-level representation in the form of contour. As mentioned in §2.5.3, contour level representation is found to be the best in cognitive musicology for melody recognition as people rely on this information to retrieve melodic information[6]. Since the work here aims to synthesize music similar to the examples presented to the system, we find that working with the contour is a logical and feasible solution to address the problem of music synthesis. MIDI format has been chosen to represent the examples and the melodic contour is formed out of the MIDI pitch(explained in section §4.4).

2.7.3 AI and Music Composition

1. Rule-Based Systems(Ebcioglü and Maxwell[28])

   The composition of music based on the set of rules. It is difficult to capture all the

rules from different styles of music. The variation in music composition is also very less as they are rule bound.

2. Model based system(Barucha and Todd [28])

   Connectionist models based on neural networks were constructed. These require a lot of training on different samples of music with respect to different parameters like pitch class, chords etc. Without accurate training the results are not pleasing.

3. Genetic algorithms(Biles [9] and Burton et al [5])

   This requires the rule representation for pitch etc to form the search domain. Rules need to be accurately represented as constraints, the fitness function need to be evaluated correctly by a human interactively. This is a tedious and time consuming job.

4. Musical Style Modelling(S.Dubnov et al[17])

   The model utilizes two theories, one is the incremental parsing(IP) and prediction suffix trees(PST). This model works at the level of motifs or patterns which are analyzed from the music files. The ability of the system to create music depends on parsing the symbols reliably for new patterns only and inserting these into dictionary. New music knowledge represented in the form of constraints can improve the performance of the system. This again depends on training the model for a variety of styles and proper encapsulation of performers knowledge in terms on constraints.

The methods and models described above have their limitations. The approach of using analogy and examples is recent in the field of music composition and has been experimented with here. Instead of codifying the rules and training a model with them, examples which inherently capture the rules can be used. This also controls the space in which the musical

elements can be varied and hence limits non-musical parameters getting in the way of generating music. Just as a composer would not like to explain each of his works in the form of rules, an example-driven approach is more intuitive and can overcome the problem of storing and analyzing too many rules.

CHAPTER 3 THEORY AND APPROACH

3.1 Extraction of video features

The perceptual features of video are extracted from the low level features. I shall give a brief overview of the method and algorithms used for low-level feature extraction. This is described in more detail in [49]. The video features that are used as input parameters in this research for mapping to audio features are color attributes such as hue, brightness, saturation, and motion.

1. HSV color space: Colors are manipulated in HSV space rather than in the RGB space for most of the video processing operations. In the HSV color cone, hues are represented by angle for each color in relation to the 0^0 line(red color). Saturation is the distance from the center of the circle. Brightness given by the vertical position along the cone. To calculate the HSV value in the frame of shot sequence

   (a) Normalize R,G,B values to[0,1].

   (b) Set max=maximum(r,g,b), min= minimum(r,g,b).

   (c) Brightness is calculated as $V_{Brightness} = max$.

   (d) Saturation

$$S_{saturation} = \begin{cases} 0 & \text{when} \quad max = 0; \\ (max - min)/max & \text{when} \quad max \neq 0; \end{cases}$$

(e) Hue

$$H_{hue} = \begin{cases} 0 & when \quad s = 0; \\ (g-b)/6.(max-min) & when \quad r = max, s \neq 0; \\ (2+b-r)/6.(max-min) & when \quad g = max; \\ (4+r-g)/6.(max-min) & when \quad b = max, s \neq 0; \end{cases}$$

if $H_{hue} < 0$, normalize it to [0-1] by $H = H + 1$;

2. Motion: To determine the motion in a sequence of frames, we need to find the magnitude of motion vector which depends on the object speed. This value is high for high object motion and low for low object motion. The work in [23] is used to compute the descriptors of motion activity. All descriptors are computed for a P frame of shot and averaged. There are several descriptors of motion for a shot. Different descriptors represent different characteristics of the video. Some of the descriptors are E_{var}-variance of motion vector magnitude, E_{median}-median of motion vector magnitude, E_{max2}-maximum motion vector magnitudes after discarding top 10% in the frame. Max2 is chosen as the descriptor which represents motion vector magnitude after removing all spurious vectors of small object movement and is found to be appropriate for home videos. It is also claimed as the best motion descriptor in [23].

3.2 Analogical Approach

3.2.1 What is the analogical approach?

Analogy is used in learning and reasoning in various situations. The use of analogies makes difficult concepts easy to understand such as comparing the concept of water resistance in pipes to the flow of electric current in a resistor. This makes generalization of concepts in one domain possible and leads to the formation of laws and then these laws can be applied to a different but similar problem in a another domain. " Learning takes place when analogy is used to generate a constraint description in one domain given a constraint in another, as when we learn Ohm's law by way of knowledge about water pipes" (Winston [37]). Analogies can be represented by descriptions in different situations and then descriptions are embodied by constraints. According to Winston, there are five key ingredients which go into making a system built on analogical learning and reasoning. Of these five principles, the principles of 'importance dominated matching' and 'classification-exploiting hypotheses' are important from this research point of view. The principle of 'importance dominated matching' lies in finding the similarity between two situations. These situations are represented by constraint relations and the best possible match is found. Matching occurs between properties, classes, acts of the situation that are similar. In order to do matching, three issues are considered.

1. To search all possible matches within a search space.

2. To constitute a quantum of evidence for a match. The quantum of evidence depends on the important constraints and relations between the two situations.

3. To combine quanta of evidence to formulate a measure of similarity. This may mean just counting the individual quantum of evidence for the particular task.

In the principle of 'classification exploiting hypotheses', memory is searched for situations that are similar to the given situation. In most of the cases, it is impractical to try all the possible pairing between the parts if the number of parts is very huge in number. So, the number of comparisons or number of possible pairings should be reduced to make it practicable.

3.2.2 Applications of analogy

The method of analogies has been earlier experimented with in image analogies [3] and in texture synthesis. The problem of image analogies can in general be defined as

$$A : A' \quad :: \quad B : B' \tag{1}$$

Here, A and A' are unfiltered and filtered version of the source images, B is the target image unfiltered version and B' is the filtered version of the target image obtained by synthesis. The new image B' is generated such that it relates to or is analogous to B in the same way A' relates to A. In texture synthesis, a model of the low level statistics of the source image is created from the filtered version and applied to a different target image. It is possible to create a super -resolution image synthesized from a low resolution one using this technique. An interesting application of analogies is the automatic synthesis of artistic styles through techniques such as example -based rendering which makes use of examples rather than rules for synthesis. Curve analogies is another application of analogies which is similar to image analogies. New 2D curves in a new different style are synthesized from examples curves in a given style. It can also be used to generate curves by modifying

existing curves to have the same style as the example [4]. Curves analogies has application in line-drawn animation and 3 dimensional(3D) signal processing.

In [45], the expressive shapes of tempo and dynamics are learnt as polynomial functions. The training phase constitutes learning the tempo variations and dynamics curves at various level of the phrase structure hierarchy. The prediction is done through using the learned rules and a simple prediction scheme based on k-NN algorithm on the partial note-level model. Though this research seems interesting, the performance has not been very satisfactory and needs to be improved by learning the local nuances which the phrase level polynomials do not capture.

In [29], there is another interesting application of the principle of analogy. This paper has explores the problem of generating motion in different styles by learning certain stylistic degrees of freedom between motion sequences(choreography). A model is trained on these parameters obtained from a data set of examples of motion cycles. Then novel motion is synthesized by interpolation or extrapolation. Analogies have been proposed as a way of bringing about extrapolation. Citing an example from the same paper, given an analogical problem defined as walking:running::strutting:X, we need to find out X. The different style co-ordinates of running, walking, strutting are used for learning. A motion sequence generated from strutting by following analogy is something which has high energy and is fast moving.

3.2.3 Application of analogies to music synthesis

A method based on the analogical method of generating images has been experimented with in generating music here. Equation 1 can be modified to suit the music synthesis method.

$$A' : A \quad :: \quad B' : B. \tag{2}$$

In this modified equation, the filtered and unfiltered versions of the example(source) and target are inter-changed. It is described as below.

A' - Filtered form of the example. A - Unfiltered form of the example music. B' - Filtered form of the new music to be synthesized. B - Unfiltered form of the new synthesized music.

In the application of this method for pitch synthesis, A' corresponds to the filtered pitch profile or contour of the example music sample, A is the pitch profile example music in the unfiltered form. The filtered version of the original is obtained from the haar wavelet approximation of it. B' is the filtered pitch profile obtained from the video features, B is the unfiltered pitch profile of the new synthesized music. B is obtained from B' by Haar wavelet reconstruction. The problem is defined in greater detail in §4.4.3.

To our knowledge, the application of analogies to music is the first of its kind. The techniques used for image analogies[3] cannot be directly applied to music synthesis. The field of non-photorealistic rendering [NPR] is based on example based rendering[EBR] but is simpler in approach than EBR and hence the properties of complex non-linear filters can be learnt quickly. NPR can be used to produce a wide range of effects while EBR is limited in its application to production of one artistic style. The idea behind [3] is to simplify image transformations by providing right example so that the desired artistic effect can be produced.

We have attempted to formulate the problem music synthesis as a filter learning prob-

lem. For this purpose, the theory of analogical reasoning is applied to the generation of music elements. Earlier models of music synthesis required large scale training of neural networks or the analysis of western classical rules which have not been very successful in generating music of one particular style §2.6. For instance, to design NPR filters for curve synthesis [4], the low-level statistics of the filter are estimated based on the position of strokes of curves. In music synthesis, low-level statistics are derived of music elements such as pitch, loudness and tempo of notes. In the framework of image analogies, there was a point to point correspondence of pixel data in the training pair. But in music synthesis a point to point correspondence cannot be so easily established. The point to point of correspondence would have to be calculated for every element of music in the training pair and applied to music sample to convert it into music with the desired effect. The greater the low-level statistics and the number of music elements considered, more the resemblance to style of the original example.

3.3 Summary

This chapter discusses the procedure of video feature extraction, the analogical approach and the way in which it can be applied to music composition. Analogy can be used for learning and reasoning in any domain. It involves the use of examples to learn the different properties of one situation and application to a different situation probably in a different domain. In the case of music synthesis task for which analogical method is used, examples of music samples are provided. The pitch space of the music samples forms the learning space. The main task is to generate musical pitch sequences from the raw pitch profile obtained from the video. The pitch space so formed by the examples is searched to find the best match.

CHAPTER 4 IMPLEMENTATION

This chapter discusses the system architecture employed for the purpose of music synthesis. It explains the calculation of audio features and the mapping of audio features to corresponding video features. The method of sequence based matching for generating the pitch of music is used here and is implemented through dynamic programming.

4.1 System Architecture

The system architecture is divided into two layers, the sonification and aesthetics layer as shown in Figure 1. These layers are explained in greater detail below.

4.2 Sonification Layer

Video features like hue, saturation, tempo are mapped to corresponding audio features. The mapping is done according to table given by Zettl in Figure 2. The sonification rules are derived from this structural mapping. The audio features that have been used are the pitch, dynamics and tempo.

4.2.1 Calculation of audio features

Pitch: The pitch values in MIDI ranges from [1 ...127]. The midi pitch key of the middle C is 60 corresponding to the frequency of 261.625Hz. The difference in frequency of one midi pitch to the next higher one is the twelfth root of 2. The range $\boldsymbol{R}$ of midi pitch in hertz is [8.1558 -12543.8539] correct to the fourth decimal place. So if the $\boldsymbol{h}$ is the normalized value of the hue of a frame, it is converted to a hertz value by $P_{hertz} = h * R + P_{min}$,where P_{min} is 8.1558 and P_{max} is 12543.8539Hz. This 'video' pitch frequency is converted into Midi note by a formula used in speech research [22]. Midi pitch note is given as

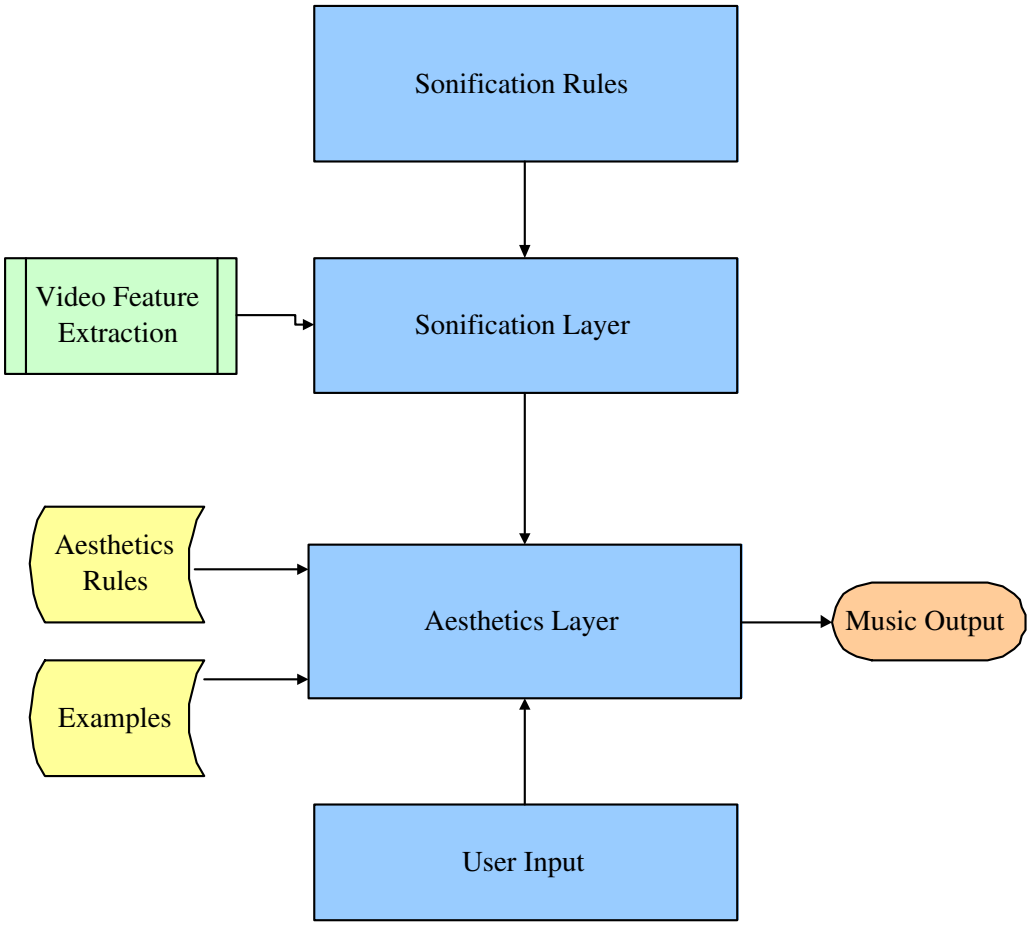

Figure 2: System Architecture

$$P_h = 60 + (ln(P_{hertz} - ln(261.625))/ln(\sqrt[12]{2}) \quad (3)$$

The table above shows the structural mapping of a portion of audio to video elements which is adapted from Zettl[50].

Audio Feature	Video Feature
Dynamics	Light Falloff Color Energy Brightness Screen Size Zooms Aesthetic Energy Magnitude
Pitch	Hue
Rhythm Timber	Light Directionality Saturation
Chords Chords & Beats Chord Tension Harmonic Complexity Harmonic Density	Texture Graphic Weight Object Placement Field Complexity Field Density
Melodic Density Contrapuntal Density Melodic Progression	Field Density Field Complexity Temporal Continuity

Table 1: Audio/Video Structural Mappping

Volume: The volume (also known as midi velocity) in MIDI ranges [0-127]. This is again derived from the brightness of the video. The conversion from brightness to midi velocity is linear. If $\boldsymbol{v}$ is the brightness of the frame, the midi velocity $\mathbf{V}$

$$V = v * 127 \tag{4}$$

Tempo: Tempo of the music generated is derived from motion and is specified in beats per minute. The duration of the notes and the motion parameter are inversely related. $\mathbf{T} = \alpha/\mathbf{m}$; α is a constant, $\mathbf{m}$ is the motion. The duration of each note and the rests in the original music is scaled according to the factor $\mathbf{T}$ which is inverse to $\mathbf{m}$. So for videos with faster motion, the note durations and rests are smaller than the shots for which the motion is slow. This ensures that the tempo of the music follows the motion in the shots and changes according to it thus synchronizing the motion with the speed of notes being played. This can be likened to a musician who varies the tempo of his composition expressively. Musicians rarely play the performance mechanically i.e. They speed up at certain places, slow down at some, stress on certain notes etc. To a pianist the important parameter dimensions are tempo and dynamics (loudness variations) which he can change continuously.

There are different tempo scales [41]. Just like pitch scales, tempo can also be arranged from lowest to highest or highest to lowest. The different tempo scales are tempo-term scales, metronome scale, harmonic scale, equal tempered scale, augmentation/diminution scale. Metronome scale was invented by Maezel after he created a device to measure tempi in terms of beats per minute. The equal tempered scale is adopted here as it similar in structure to the pitch scale used in the composition of serial music. The tempo scale is

Equal Tempered scale MM - (metronome marking) whole note
60
63.6
67.4
71.4
75.6
80.1
84.9
95.2
100.9
106.9
113.3
120

Table 2: Equal tempered scale

divided into 12 tempi per octave and the ratio between successive tempi is the twelfth root of two. The shortest duration in the scale is the sixteenth note at MM (metronome marking) 120 i.e. one thirty second of a second, the longest note duration is of octal whole note at MM whole note 60 i.e. eight seconds. The equal tempered scale is as given below and is used for research here.

4.3 Aesthetics Layer

The aesthetics layer is a compositional layer that gives form and structure to music generated by the lower level sonification layer. The output from the sonification layer is sound which is not musically pleasing. A note is generated for every frame and lasts for the duration of the frame. To be able to compose music requires much more than the raw data obtained from the sonification layer. So, the aesthetics layer contains this knowledge of

musically shaping the output further. Music can either be generated through a rule based approach where rules are used to train a system or, by means of an example based approach where examples of music are used as reference to generate it. The rule based approach as pointed out in the literature review has the disadvantage of codifying every rule which requires extensive music knowledge besides being very painstaking. So, the research here adopts another approach which is quite recently used in image synthesis methods for generation of different styles but has not seen any precedents in computer music composition. The approach of using examples for the music synthesis is novel.

After obtaining the pitch of synthesized music by matching with that of example, the midi velocity and tempo is also calculated. These parameters values which are in the form of vectors are then assembled together in the text format of MIDI. This MIDI text file is then converted to MIDI music through the MIDI IO library.

4.4 Analogy based Composition

Using the analogy based method, we generate the pitch of the new music by matching the pitch profile (contour) derived from the video with the pitch contour of the example chosen.

4.4.1 Sequence based pitch matching

The sequence -based comparison method is used widely in the realm of molecular biology, applied to geological data, speech research etc. When we compare any two sequences of data, then can be found to be different in the following ways.

1) Substitutions (also known as replacements) 2) Deletions and insertions (indels) 3) Compression and expansion 4) Transposition Most of the sequences that occur in real-life applications exhibit the substitution difference though other differences such as deletions,

insertions are also important. The difference between sequences can be analyzed in three ways. a) A trace b) Alignment c) Listing.

To find a trace of two sequences a and b, the elements that are common to both the sequences are taken first in the order found in the sequences. In addition to these elements, the elements that are in the target sequence are added and the ones in the source sequence without a corresponding match in the target are deleted. Alignment is similar to trace but the elements considered for alignment need not be in order. A listing is an algorithm invoked by the sequence of edit operations such as insertions and deletions.

Sequence comparison can be done through alignments, traces, listings. In applications for determining whether the molecules are homologous or not, the distance measure is important. The distance measure depends on the application and in applications such as continuous speech recognition, the distance measure is used to find an optimal alignment.

1. *Distance for sequence comparison.*

   The meaning of distance has different definitions. Levenstein introduced two distance measures: one based on insertion-deletion and the other based on substitutions [40]. The definition of distance can be built on the following points.

   (a) Elementary operations are substitutions and insertions, deletions.

   (b) All listings from sequence a to sequence b are based on these operations.

   (c) The length of the listing is based on the number of elementary operations which are mentioned above.

   (d) Distance between the sequences is minimum length of the listing.

   The length of any listing is given by number of indels + w (number of listings) [40]

The weight for every operation is given by w. If $w > 2$, then a deletion-insertion pair is used in place of an insertion and $w = 2$ implies that an insertion and deletion is equal to a substitution. The other distance for comparing sequences a and b are Euclidean distance $\sqrt{\sum(a_i - b_i)^2}$, City-block distance $\sqrt{\sum(a_i - b_i)}$ and hamming distance which refers to the number of positions in which the two sequences differ. But there are situations in sequences comparison when a and b are not equal in length or the sequences are very different from each other that the correspondences between a and b is not known and these distances cannot be used. In order to find the optimal alignment between the two sequences, the preferred method is to find every possible correspondence between a and b and then find optimal alignment between them. The procedure of finding the alignment falls under the framework of dynamic programming.

2. *Edit distance* Edit distance is defined as the minimum number of local transformations required to transform a sequence A $=a_1, a_2, ..., a_n$ into another sequence B= $b_1, b_2, ..., b_n$ and denoted by *D(A,B)*. The local transformations are

   (a) replacement $a_i -> b_j$.

   (b) insertion $\lambda -> b_j$.

   (c) Deletion $a_i -> \lambda$.

Properties of the *D (A, B)* metric.

(a) $D(A, B) >= 0; (D(A, B) = 0 iff A = B)$

(b) $D(A, B) = D(B, A)$

(c) $D(A, C) <= D(A, B) + D(B, C)$.

4.4.2 Notation

A finite set of symbols called the alphabet.

$\Sigma \epsilon 1, 2, ..., 127$	Alphabet of Midi pitch notes.
A $= a_1, a_2, .., a_m$	Sequence of over $\Sigma; a_i \epsilon \Sigma$.
$\|A\| = m$	The length of A
$P_m = p_{m1}, p_{m2}, ..., p_{mn}$	Sequence of midi pitch notes of music sample
P_{mh}	Haar approximation of the music sample. Level of approximation is 5.
$\|P_{mh}\|$ = m	Length of midi pitch sequence P_{mh}
$h(t)$ ϵ $0, .., 1$	Hue profile of the video w.r.t to time.
P_h ϵ $1, 2, .., 127$	Pitch profile derived from the video $h(t)$.
P_h	Pitch profile derived from video
P_{hh}	Haar approximation of pitch profile of the video
$\|P_{hh}\| = n$	Length of the Haar approximated pitch profile P_{hh}
T_m	Tempo of music scale
M_v	Motion of each shot of video
T_v	Tempo derived from video
P_{mat}	Midi pitch sequence after sequence matching
P_{syn}	New Sequence of synthesized pitch

4.4.3 Problem Definition

Given the midi pitch profile of music sample P_m and its approximation P_{mh}, video pitch P_h and its approximation P_{hh}, obtain the new synthesized midi pitch profile. Given the tempo of the video T_m and the motion M_v of the video, obtain the tempo of the music that depends on the motion of the video.

Writing the pitch composition problem in terms of analogy,

$$P_{mh} : P_m \quad :: \quad P_{hh} : P_{syn} \tag{5}$$

The aim is to P_{syn} given the other three terms in the analogy.

4.4.4 Algorithm for music composition

The above diagram shows the algorithm for music composition which is explained in the subsequent sections.

4.4.5 Sequence based comparison

1. Procedure to obtain P_{syn} through analogy

   Consider P_{mh} and P_{hh} as two sequences of pitch midi notes. Alternatively we can consider these sequences to represent the contours of the pitch sequences in the filtered form. To obtain the new set of synthesized notes, the matching between the contours is done using sequence comparison techniques so that the synthesized music emulates a particular example. The lower resolution form of both the music sample and that of the pitch profile is considered. This is obtained by taking the Haar approximation of these profiles.

   To measure similarity between two sequences of pitch segments, P_{mh} (the Haar approximate music pitch profile) and P_{hh}(the pitch profile from video), we need to calculate the local transformations which are replacement, insertion, deletion. The sequence can be obtained from and by a set of transformations steps $a_1 -> b_1, a_2 -> b_2, .., a_n -> b_n$.

2. Evaluation of the edit distance through Dynamic Programming

   P_{hh} and P_{mh} are considered to be row vectors. The edit distance is the minimum number of local transformations that are required to transform into and this can be calculated using dynamic programming. The procedure consists in constructing an integer matrix where each row corresponds to the event(note) in P_{hh} and each column to that of P_{mh}. Each cell stores gives the element similarity distance $d_{i,j}$ between

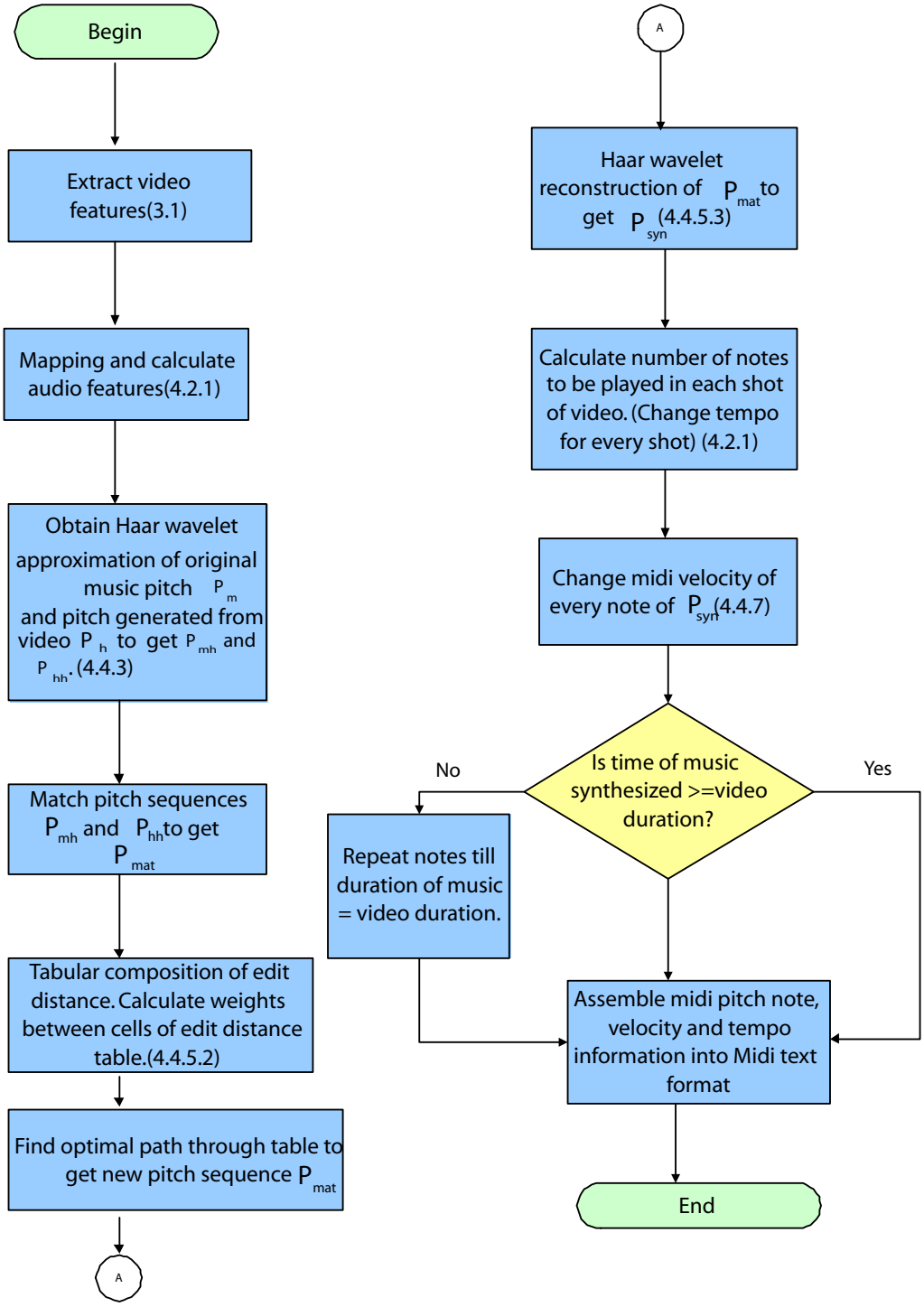

Figure 3: Music Composition Algorithm

$a_i$and b_j where $i \epsilon 1, 2, ..., n$ and $j \epsilon 1, 2, .., m$. Element similarity distance

$$d_{i,j} = \frac{|(a_i - b_j)|}{(a_i + b_j)} \tag{6}$$

There are two types of edit distances. Operation weights edit distances which depend on the edit operation like insertion, deletions and substitutions. Each of these operations has their attached weights. The substitution is considered to be a deletion followed by an insertion. The objective function of an operation weight distance is to minimize the total weight of the edit operations. The other edit distance is alphabet weight edit distance where weight or score of a substitution depends on the characters of the alphabet and not on the position of the character in the string. $d_{i,j}$ is the distance based on characters itself and not on the position of the characters in the string. Hence this would represent the alphabet weight edit distance which is used to calculate the weighted edit distance graph.

The sequence matching between P_{hh} and P_{mh} can be efficiently done through dynamic programming. Dynamic programming has three components a) Recurrence relation b) Tabular computation c) Traceback.

(a) *Recurrence relation*: The calculation of edit distance is the based on finding $d_{i,j}$ recursively for all i and j. Instead of a recursive function, which is computationally inefficient though easy to program, a tabular computation of the values of P_m and P_h is considered.

(b) *Tabular computation of edit distance*. This is a bottom up approach of computing the edit distance. It is done by laying out the dynamic programming table of size $m * n$. Cells of the table holds the value of $d(i, j)$ for all values i and j.

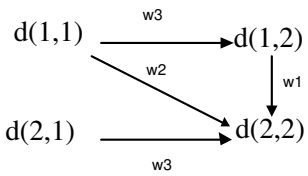

Figure 4: Weighted arcs between cells

The table is filled one row at a time. In order of increasing i and within each row in order of increasing j.

(c) *Traceback*. We need to find the optimal edit transcript which is the optimal patch out of the several paths from the source (0,0) to destination (m,n) to get the optimal alignment between the two sequences of pitch notes. In the edit transcript, each horizontal edge from cell (i,j) to cell (i,j-1) is interpreted as an insertion (I) of character $P_{mh(j)}$ into P_{hh}, each vertical edge from (i,j) to (i-1,j) as a deletion of $P_{mh(i)}$ from P_{mh}. And each diagonal edge from (i,j) to (i-1,j-1) as a match(M) if $P_{hh(i)} = P_{mh(j)}$ or a substitution(R) if $P_{hh(i)} = P_{mh(j)}$.

To find the optimal path we need to construct a weighted edit graph, the key property of such a edit graph is that any shortest path (the total weight is minimum) from (0,0) to (m,n) specifies a edit transcript with minimum number of edit operations. We can consider each $d(i,j)$ as a node in the graph. Each $d(i,j)$ is linked by arcs as shown in Figure 4 where $-R <= i - j <= R$ where R is the band of matching window . Here R is one. Three directed arcs are used to link three vectors with d *(i,j)* with weight w2, $d(i-1,j)$ with w1 and $d(i-1,j-1)$ with w2. The weights stand for the cost of edit operations such as substitution (w2), insertion (w3), deletion (w1), the cost being zero for a match.

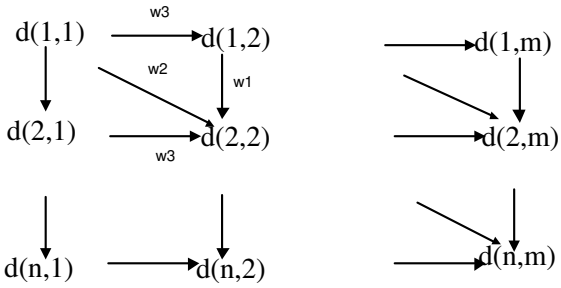

Figure 5: A directed graph from similarity distance matrix

A directed graph is constructed from this similarity distance matrix as given in 5 where $d_{1,1}$ the initial vertex is and $d_{n,m}$ is the terminal one. Weights along the arcs of the directed graph are given as

$$w_1 = d_{i,j} \tag{7}$$

$$w_2 = 2 * d_{i,j} \tag{8}$$

$$w_3 = d_{i,j} \tag{9}$$

$$\tag{10}$$

The weight of a substitution is double that of insertion and deletion because the table is not square with different number of rows and column. This ensures that the optimal path is not slanted towards the 45^0 line or the diagonal of the edit distance table which will only be correct for a square table.

(d) Procedure to find distance between the two pitch series

For each note segment in P_{hh}

Begin

For each note segment in P_{mh}

Begin

$d_{i,j} = |(a_i - b_j)|$ \($a_i + b_j$)

End

End

The optimal path is found through the shortest path from D(1,1) to D(n,m) is calculated by an efficient graph matching algorithm. Here Djikstra's algorithm is used for fast computation and gives new pitch profile representing the new set of synthesized notes. This gives the filtered form P_{new} of the new music to be generated.

To get the actual synthesized music, the filtered form P_{new} is taken to be haar approximation of the new music. A Haar wavelet reconstruction is then done by taking the detailed co-efficients of the original music sample P_m and approximate co-efficients of P_{new} to get synthesized midi pitch sequence P_{syn}.

4.4.6 Midi velocity of synthesized music

The midi velocity is obtained as described in §4.2.1. The brightness of the frames is averaged over a shot. The maximum and minimum velocity in the sample music is calculated. The velocity of the notes in P_{syn} is linearly varied over every shot by modifying (Eqn: 4). If V_{min} and V_{max} are the minimum and maximum velocity of the sample music, then the velocity of the synthesized notes over a shot is $V_{note} = V_{min} + 127 * ((V_{max} - V_{min})$

4.4.7 Complexity of computation

The time analysis of dynamic programming method of calculating edit distance results in *O (nm)*. That is constructing the dynamic programming table of computing edit distance

D (n, m) between a pitch sequence of length n and a sequence of length m is filled in *O (nm)*. Each cell takes constant number of calculations. Generating the weights to construct the weighted graph also takes constant time because it involves constant number of examinations, operations. Each cell is connected to three other cells (i-1, j-1), (i, j-1), (i-1, j) by means of weights *w1, w2, w3* (Eqn:8,9,10). The total time for calculating the weights is $O((2*(m-1)(n-1)+k))$ where $k = min\ (m,\ n)$. The optimal edit transcript can be computed in O $((V+E)logV)$ using Djikstra's algorithm, where V is the number of vertices (each cell) in the table, E is the number of edges. So, the total computational time is

$$O(nm) + O((2*(m-1)(n-1)+k)) + O((V+E)logV);$$
$$max(V) = m+n, V = E, k = min(m,n) \tag{11}$$

4.5 Summary

This chapter has explained the system architecture used in the composition of music. The sonification layer converts the video features into audio parameter values directly. The aesthetics layer adds the musical knowledge to make the music generated pleasing. This is done by providing examples to that layer. The pitch sequence from the sonification layer is matched with that of the music sample by applying the method of dynamic programming used in sequence based matching. The tempo variation is directly related to the motion of the video shots. The loudness is also varied according to the brightness of the shots. These MIDI parameter values are then assembled into the MIDI text format and then converted into MIDI music by means of MIDI IO library. Though the sonification layer automatically outputs the audio parameter values, the aesthetics layer involves the user. The user can

select the example music of his/her choice and generate new music.

CHAPTER 5 EXPERIMENTS AND RESULTS

5.1 Implementation Platform

The music composition for every video has been implemented on the Microsoft windows platform. Microsoft Visual c++ has been used for implementing the video feature extraction code. MATLAB has been used to implement the code pertaining to sequence comparison algorithm. MIDI IO library has been used to convert midi music files into midi text format and midi text format into midi music.

Configuration of computer system

CPU: Intel Pentium 111 866 Hz.

Memory: 132MB RAM.

Storage: 10 GB

OS: Windows NT 2000 professional

5.2 Procedure

The music data in our experiments are melodies mainly selected from western classical instrumental music. There are midi archive websites dedicated to western classical music from where the melodies have been collected [1]. The midi files have to be first converted to text files by MIDI IO library [2]. The test file gives onset time, duration of each note, midi pitch note and velocity of every note.

The test videos are run through the DVA system to get the color, motion information. The hue is converted to midi pitch notes and the midi pitch profile is obtained as shown in Figure 6. The motion information is similarly obtained and then converted to tempo number based on the even tempered scale.

Video	Description of video	User selected Melody	Tempo(in BPM) Synthesized Music	Hue-pitch profile
Airplane	*Motion*:rises from slow to fast, stabilizes at medium pace and ends in slow pace *Hue*: Changes only slightly between frames. (a blue background throughout video). *Brightness*: almost even throughout the video.	Dmajor Dminor Bbmajor Emajor Gminor Cmajor Well tempered Clavier Cminor	2- 120- 72- 120- 108- 60- 84- 84- 84 - 84 - 84	Figure 7
Motorola	*Motion*:Almost a medium paced motion which, increase towards the end of video *Hue*: Slight changes present. between frames. *Brightness*: Low Brightness. Increases in middle shots.	Dmajor Bbmajor Emajor Cmajor Well tempered Clavier Cminor Gminor	120- 120- 60- 120 - 120	Figure 11
Tango	*Motion*: An average paced motion with bursts, of high energy motion in some shots. *Hue*: Slight changes present between frames. *Brightness*: Low Brightness throughout the video.	Emajor GMinor Emajor	120- 120- 60- 120 - 120	Similar to Figure 11

Table 3: MPEG videos used for survey

5.3 Results

The results were tested to find out whether the music composed met the satisfaction of the users. The results were evaluated on the following criteria:

1. Synthesized quality of music: This refers to the how pleasing the music sounds to the ear.

2. Aesthetic appeal of music: This refers to whether the music enhances the video appeal.

3. Appropriateness of music for the video clip: This refers to whether the music seems to fit the content and semantics of the video or detracts from it.

4. Tempo of music: This refers to whether the tempo in the music follows the motion in the video.

The scale ranged : 1-Unacceptable, 2-Needs Improvement, 3-acceptable, 4-Good, 5-excellent. The survey results for each MPEG videos used and the overall feedback on the synthesized music is given below:

.

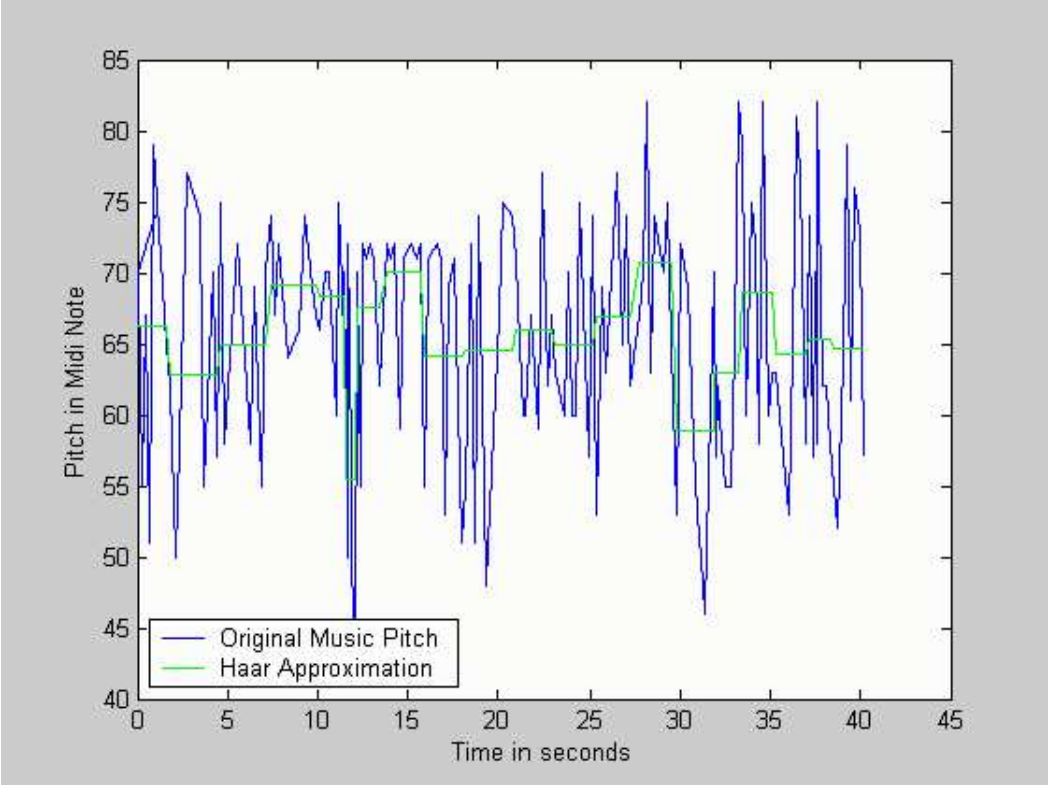

Figure 6: Pitch Contour of Gminor Bach melody and its Haar Approximation

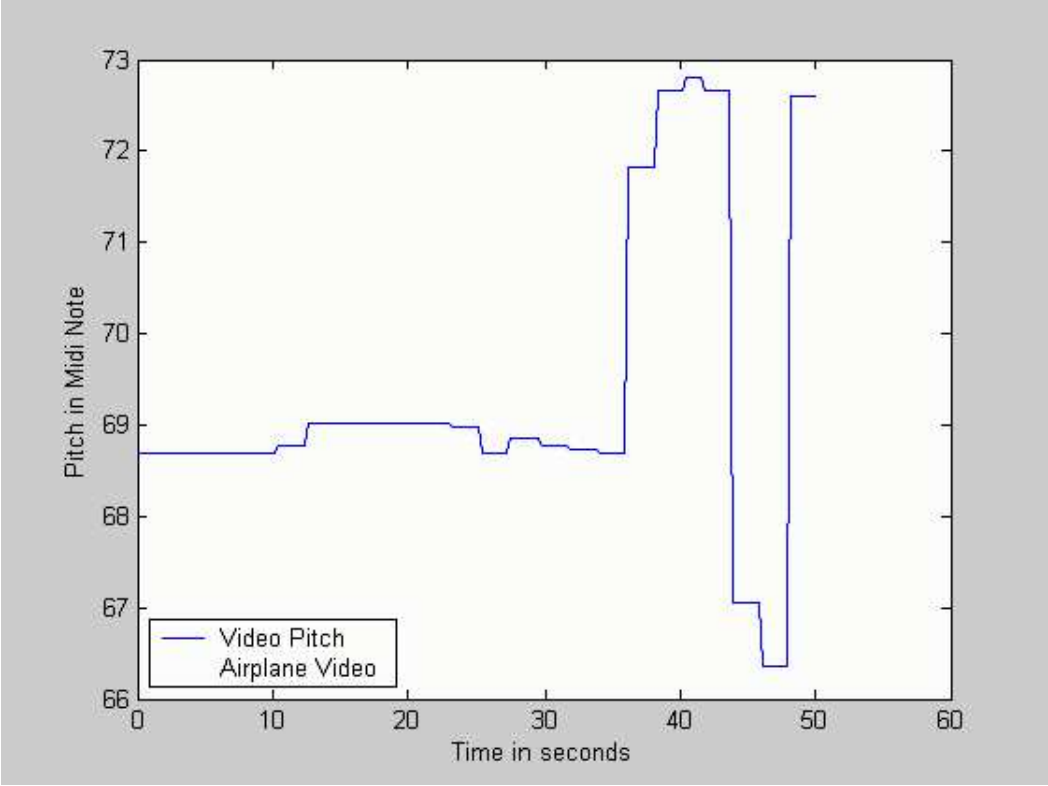

Figure 7: Pitch Contour of 'Airplane' video clip obtained from sonification layer.

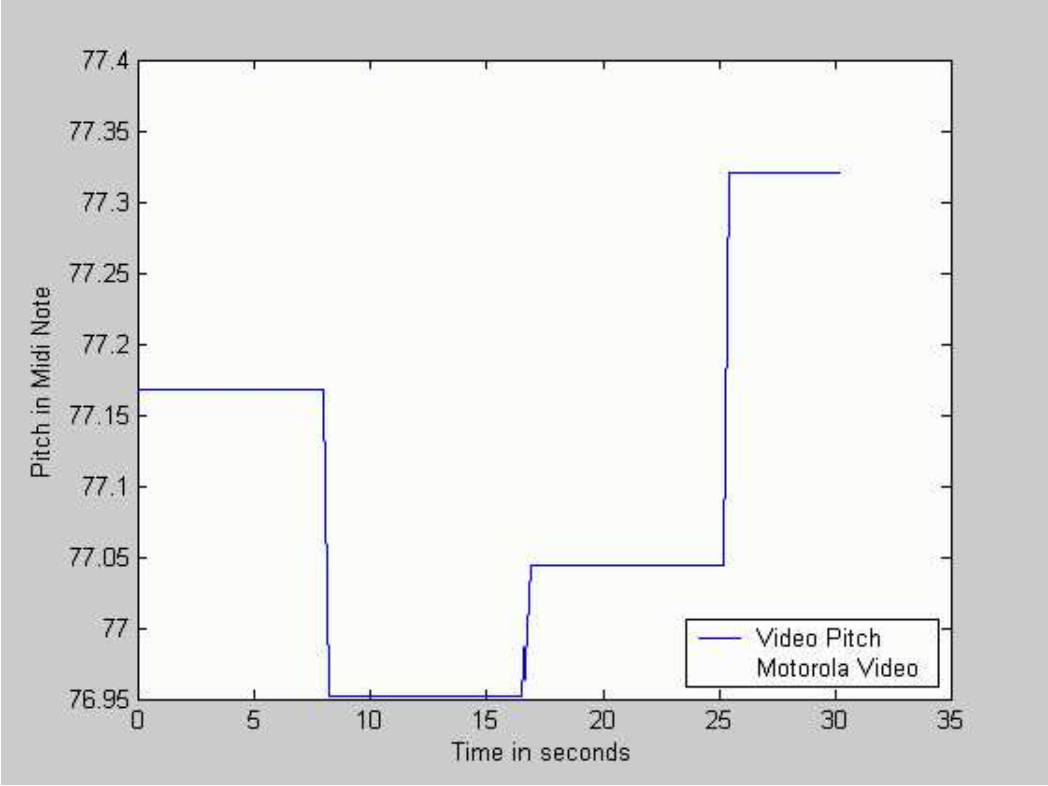

Figure 8: Pitch Contour of 'Motorola' video clip obtained from sonification layer.

Video	Usability Criteria	Rating of Synthesized Music
Airplane	Is the Music pleasing?	3.1
	Is the music appropriate to video clip?	2.8
	How well does tempo match the pace of the activity in video?	3.14
Tango	Is the Music pleasing?	2.83
	Is the music appropriate to video clip?	2.3
	How well does tempo match the pace of the activity in video?	2.65

Table 4: Survey results on synthesized music for MPEG videos

How well does the music enhance video appeal?	3.2
Your satisfaction with synthesized music?	2.8
Is the instrument used in music chosen correctly?	2.2

Table 5: Survey results on overall quality of synthesized music

Figure 6 shows the pitch profile of the Bach melody and its Haar Approximation. The dotted line in red is the Haar approximation and the solid blue line is the actual pitch in MIDI. Figure 8 shows the pitch from the video which is derived from the hue of the video frames. It shows that the pitch doesn't vary as much as the example music pitch. Figure 9 and Figure 10 is the pitch contour obtained after the sequence matching of hue profile of the airplane video with the example music pitch of Emajor and Gminor melodies. Figure 11 is the pitch contour obtained after the sequence matching of the hue profile of motorola video with the example music pitch of a melody in Emajor.

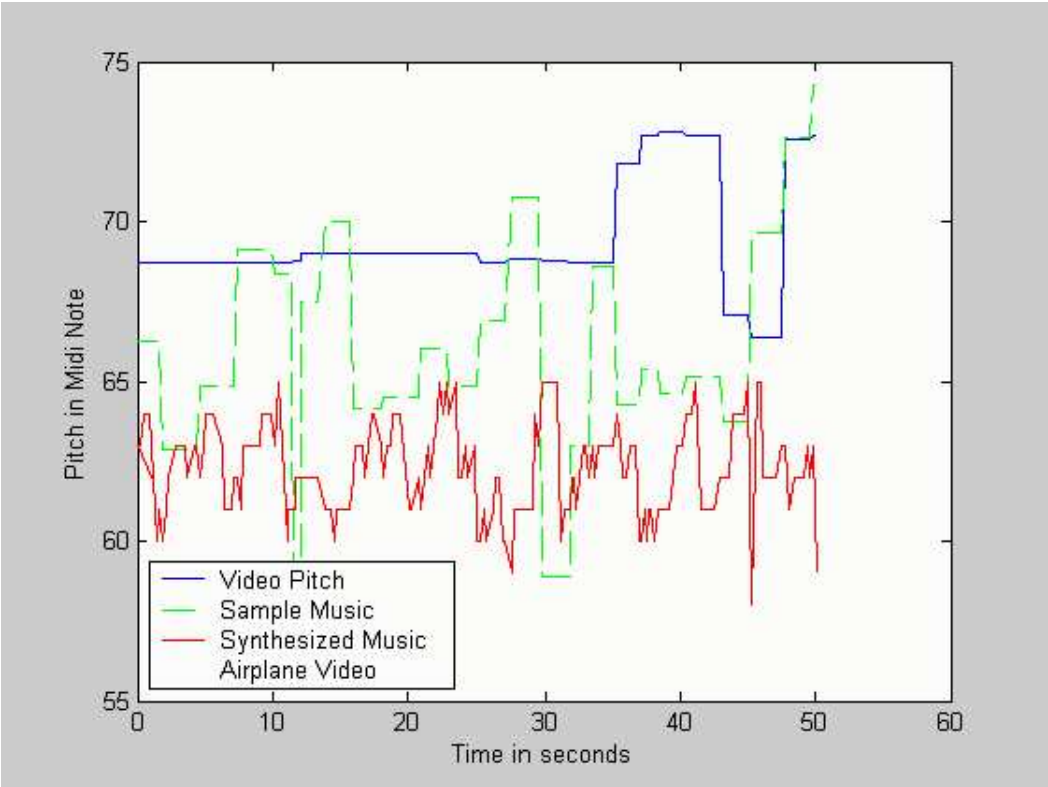

Figure 9: Pitch Contour of synthesized music with Gminor(Airplane Video)

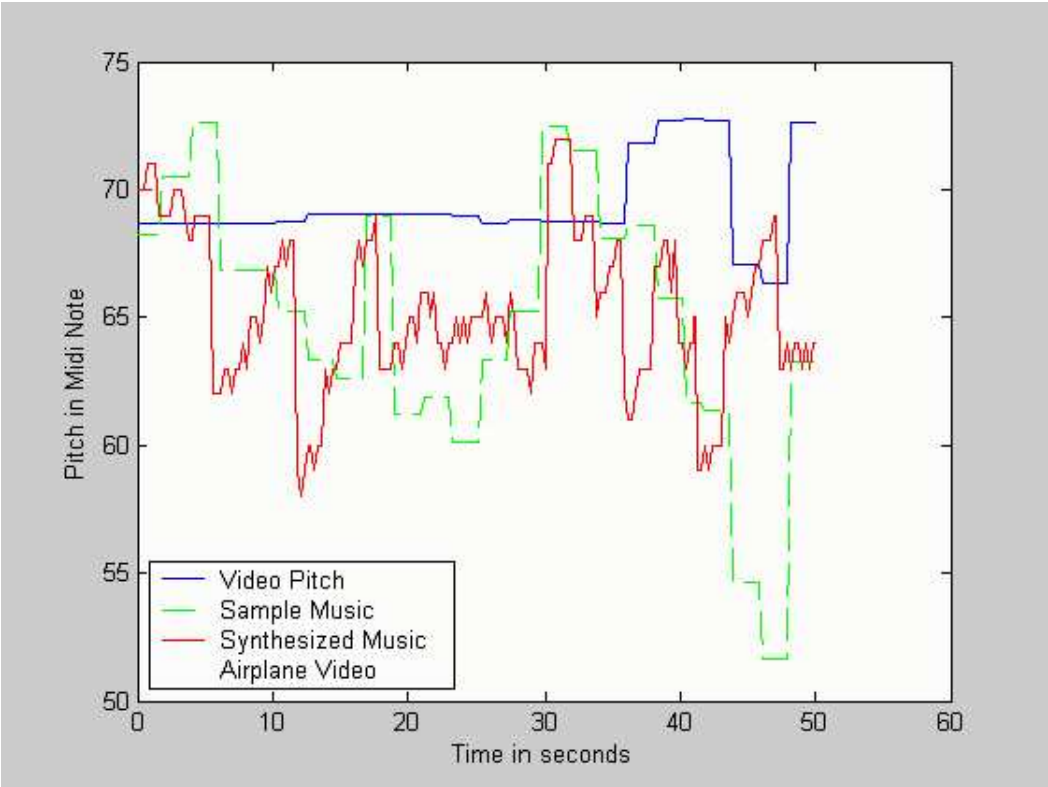

Figure 10: Pitch Contour of synthesized music with Emajor(Airplane Video)

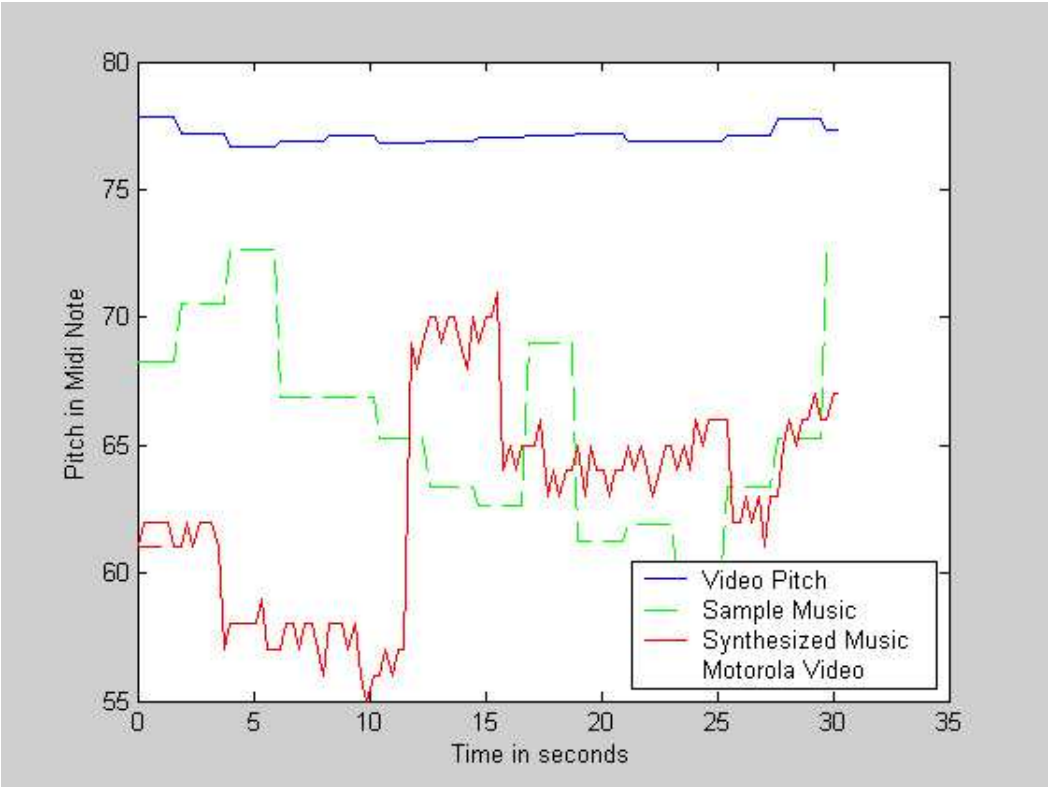

Figure 11: Pitch Contour of synthesized music with Emajor(Motorola Video)

CHAPTER 6 ANALYSIS AND CONCLUSION

6.1 Analysis

The pitch contour of the video as shown in Figure 6 is derived from the hue of every frame of the video. The sequence comparison method gives us the notes that are 'similar' to the music example. The velocity of the note is also computed from the brightness of video and assigned to every note. The pitch, volume so generated are re-assembled in the midi format and then converted to midi music using MIDI IO library. The user survey done on the analogy results suggests that the music generated is acceptable though some parts of music may not seem to be musically pleasing. The experiments have been tried with one instrument, the piano. The melodies are basically selected from the collection of western classical music. The reason behind this is two-fold. Firstly, the practical consideration that of the relative ease of finding and availability of midi files for western classical music. The other reason was that the western classical music follows the certain rules. These rules concerning the pitch intervals, timing of the notes are implicitly followed when the music contour is imitated without using a rules database. From the results, one can also deduce that tempo changes are more perceptible and tangible than the pitch changes, especially when there are motion changes through the shots from fast to slow and vice-versa . Some users have also noted that the mixing of the literal sounds (the source audio) in the video with the music generation would enhance the appeal. This in general is preferred to the rule based generation of chord music which we had experimented with earlier. Music synthesis for chords was done using other methods such as generating major and minor chords using the rule-based approach. The user survey shows that music by analogy scores over the other methods of music composition.

6.2 Discussion

In this dissertation, we have presented a novel approach to add audio to video which focuses on generating content-related music by translating primitive elements of the video to audio features and using sequence comparison to synthesize new pitch sequence. The tempo of the music is also varied according to the motion of the video sequences. So, the music tries to be in synchrony with the events happening in the video. But the area of music composition if vast and one can find ways of improving these results in many different ways by considering different parameters which enrich the music uniquely. The challenge in such a task lies in selecting the music elements that can be determined or affected by the video elements through the principles of movie production or aesthetics heuristics of audio-video mixing as we have done here.

Human beings learn to sing by imitation or through training in classical forms of music by imbibing the rules that are allowed in classical music. To enable a computer to compose artistically skillful music, it needs to either understand the rules of music generation or learn it through examples. This research has concentrated on making use of examples to provide the input data. Examples restrict the space used for searching for new combinations of pitch, dynamics and other elements. It is therefore more feasible and also more intuitive just as a composer would think of composing new music in his own style. The music synthesized was obtained by generating the pitch, tempo and dynamics matching hue, motion and brightness of video elements.

The appreciation of music also depends on personal choice and taste. Music has evolved through generations , from the western classical music culture to the acid rock of today. What may appear to be discordant notes to one may not be perceived the same by another person. The survey has indicated that the music composed is not unpleasant to the ears and that it does enhance the video appeal when compared to the silent video. But it could

be improved as mentioned in the following section. It also appears that some people prefer semantically relevant music as it appears more real and natural: as one user suggests "I would prefer the sound from the objects in the video and the music just act as the background. I meant if the focus is a plane, then the plane's sound and if the focus is the children dancing, the children's voices/laugh." This requires a high level of semantic knowledge of the video objects.

6.3 Recommendations for future work

Based on the above research, we can conclude that the system could be expanded to include more musical dimensions in order to produce richer music and semantically relevant music. Firstly, the pitch produced can be further enhanced by considering the arrangement of chords and selecting the music in a particular key such that it matches the semantic level of the video. As has been mentioned earlier, the major keys are appropriate for brighter moods and minor keys for dull, somber moods. Including this feature may require extraction of some new video features and generation of tonal music according to key.

Secondly, another parameter that could be varied is the instrument. The experiments were tried with piano music. Though most of western classical music is on the piano, one instrument for all videos may not adequately bring in the variety required and may seem boring and monotonous. So, we can try to create polyphonic music by generating music through different instruments that are again chosen by some video feature such as the saturation. This can be a hard problem as one needs to be careful in selecting instruments that can produce pleasing music when played together and doesn't sound discordant.

Lastly, a music model can be built by training examples using analogy so that music can be created in a particular style of a composer selected by the user. The system can take in the users' choice of style and generating music according to the chosen style. In

order to do this, one needs to identify the composer's signature which can be composed of the type of instruments he uses most to harmonic progressions of chords among other musical elements.

References

[1] Home page of midi archives http://www.midiarchives.com.

[2] Home page of midi io library. http://midiio.sapp.org.

[3] Hertzmann A, Jacobs C, Oliver N, Curless B, and Salesin D. Image analogies. In *Proceedings of ACMSIGGRPAH*, pages 327–334, 2001.

[4] Hertzmann A, Oliver N, Curless B, and Seitz S. Curve analogies. In *Proceedings on Eurographics Workshop on Rendering*, 2002.

[5] Burton A.R and Vladimirova T. Generation of musical sequences with genetic techniques. In *Computer Music Journal*, volume 23, No.4 of *Issue 1*, Dec 1999.

[6] Lindsay A.T. *Using Contour as a mid level representation of melody.* PhD thesis, MIT, 1996.

[7] Balaban.M, Ebcioglü.K, and Laske.O. *Understanding Music with AI: Perspectives on Music Cognition.* MIT Press, 1992.

[8] Bergman.A. *Auditory Scene Analysis.* MIT Press, 1990.

[9] Biles.J. Genjam: A genetic algorithm for generating solos. In *Proceedings of the 1994 International Computer Music Conference(ICMC)*, 1994.

[10] Burns.E and Ward.D. *Intervals, Scales and Tuning.* Psychology of Music, Academic Press, 1982.

[11] Chion.M. *Audio-Vision.* Columbia University Press, New York, 1994.

[12] Dorai.C and Venkatesh.S. Bridging the semantic gap in content management systems: computational media aesthetics. In *International Conference On Computational Semiotics in Games and New Media*, pages 94–99, 2001.

[13] Dorai.C and Venkatesh.S. *Media Computing: Computational Media Aesthetics*. Kluwer Academic Publishers, 2002.

[14] Dowling.J. Scale and contour:two components of a theory of memory for melodies. *Psychological Review*, 85(4):341–354, 1978.

[15] Dowling.J. *A Brief Survey of Music Representation Issues,Techniques, and Systems*. Psychology of Music, Academic Press, 1982.

[16] Dowling.J. *Melodic Information Processing and its development*. Academic Press, 1982.

[17] Dubnov.S, Assayag.G, and Bejerano.G. Using machine-learning methods for musical style modelling. *IEEE Computer*, October 2003.

[18] Youngmoo.K et al. Analysis of contour based representation of melody. *Proceedings of International Symposium on Music Information Retrieval*, 2000.

[19] Yuehu Liu et al. A method for content-based similarity retrieval of images using two dimensional dp matching algorithm. In *11th International conference on image analysis and processing*, September 2001.

[20] Foote.J, Cooper.M, and Girgensohn.A. Creating music videos using automatic media analysis. In *ACM Multimedia*, pages 553–560, 2002.

[21] Hiller.L and Issacson.L. *Experimental Music*. McGraw-Hill, 1959.

[22] Apple Computer Inc. *Inside Macintosh: Quicktime*. Addison-Wesley, 1993.

[23] Peker K, Divakaran A, and Papathomas T. Automatic measurement of intensity of motion activity of video segments. In *Proc. SPIE*, volume 4315, pages 341–351, 2001.

[24] Kamien.R. *Music: an appreciation*. McGraw-Hill, 2000.

[25] Lemstrom.K. *String Matching Techniques for Music Retrieval*. PhD thesis, University of Helsinki,Finland, November 1999.

[26] Lerdahl.F and Jackendoff.R. An overview of hierarchical structure in music. *Music Perception*, 1(2):229–252, 1984.

[27] Leri.D. The fechner weber principle. http://wwww.semiophysics.com.

[28] Loy and Todd. *Music and Connectionism*. MIT Press, 1991.

[29] Brand M and Hertzman A. Style machines. *Proceedings of SIGGRAPH*, pages 183–192, July 2000.

[30] Srinivasan M, Venkatesh S, and Hosie R. Qualitative extraction of camera parameters. In *Pattern Recognition*, volume 4 of *30*, pages 593–606, 1997.

[31] Mazzoni.D and Dannenberg R.B. Melody matching directly from audio. In *ISMIR*, 2001.

[32] Mulhem.P, Kankanhalli M.S, Hassan.H, and Ji Yi. Pivot vector space approach for audio-video mixing. In *IEEE Multimedia*, volume 10, No. 2, pages 28–40, Apr-Jun 2003.

[33] Narmour.E. *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. Chicago: University of Chicago Press, 1990.

[34] Home Page of DIVA project. http://diva.comp.nus.edu.sg.

[35] Home Page of MULTIQUENCE Technologies. http://wwww.goldwave.com.

[36] Home Page of Muvee Technologies. http://wwww.muvee.com.

[37] Winston P.H. Learning and reasoning by analogy. *Communications of the ACM*, December 1980.

[38] Dannenberg R.B. An on-line algorithm for real-time accompaniment. In *Proceedings of the International Computer Music Conference*, 1984.

[39] Dannenberg R.B. A brief survey of music representation issues, techniques, and systems. *Computer Music Journal*, 17 No. 2:20–30, 1993.

[40] Sankoff.D and Kruskal.J.B. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison.* Addison-Wesley Publishing Company Inc, 1983.

[41] Greschak.J. Tempo scales in polytempo music. www.greschak.com/polytempo/ptts.htm.

[42] Selfridge-Field.E. *Beyond Midi:The handbook of musical codes.* MIT Press, 1997.

[43] S.H.Srinivasan, Meera G Nayak, and Mohan S.Kankanhalli. Music synthesis for home videos. In *manuscript under preparation*, September 2003.

[44] Cormen T.H, Leiserson C.E, and Rivest R.L. *Introduction to Algorithms.* MIT Press, 1990.

[45] Tobudic.A and Widmer.G. Proceedings of the 5th international conference on case-based reasoning. Trondheim, Norway, 2003.

[46] Press W.H, Teukolsky S.A, Vetterling W.T, and Flannery B.P. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 1994.

[47] Widmer.G. The synergy of music theory and ai: Learning multi-level expressive interpretation. In *Proceedings of the Twelfth National conference on Artificial Intelligence(AAAI-94)*, pages 114–119. AAAI Press/MIT Press, 1994.

[48] Xenakis.I. *Formalized Music: Thought and mathematics in composition*. Indian University Press, 1971.

[49] Ji Yi. Video features for audio-video mixing. Master's thesis, National University of Singapore, 2003.

[50] Zettl.H. *Sight,sound,motion: Applied Media Aesthetics*. Wadsworth, 2 edition, 1998.

APPENDIX A - USER SURVEY

PART I: User Background

1. Which statement best characterizes your music background?
   (a) I am professionally trained in Music. If yes, which style?(Please specify).
      i. Western
      ii. Eastern
      iii. Others
2. I have no formal training in music but possess music knowledge. If so,
   (a) Western
   (b) Eastern
   (c) Others
3. I am a casual listener.
4. I am totally ignorant

PART II: RATING

1. 5 - Excellent
2. 4 - Good
3. 3 - Acceptable
4. 2 - Needs Improvement
5. 1 - Unacceptable

Please provide comments in addition to the rating. Please give the rating for each video clip separately. Please use the comments column.
The results are on the page http://www.comp.nus.edu.sg/ meeragaj/index.html

PART III: USER FEEDBACK

Music Analogy	Rating		Comments
	Video A	Video B	
1. Is the music pleasing? 2. Is the music appropriate with the video clip? 3. How well does tempo of music match pace of activity in video (For e.g. is the music too fast and activity of video slow?)			
Overall Rating			
1. How well does the music enhance video appeal? 2. Your satisfaction with the synthesized music 3. Is the instrument used in the music chosen correctly?			

Overall Comments/Suggestions for improvement

APPENDIX B - STANDARD MIDI FILE EXAMPLE

1. Header Information
   (a) Key signature: 0
   (b) Time Signature: 3 4
   (c) Ticks per quarter note: 96
   (d) Number of 32nds per quarter note: 8

2. Track Chunk Information

MIDI Information		Pitch Information		Duration Information		Performance Information
Track no	Event Type	Note Name	Key No	Delta Time	Elapsed Time (Bars:Beats:Ticks)	Dynamic Level (Velocity)
1	Note On	C5	72	0	01:04:000	48
1	Note off	C5	72	48	01:04:048	48
1	Note On	E5	76	0	01:04:048	48
1	Note off	E5	76	48	01:04:096	48
1	Note On	G5	79	0	02:01:000	48
1	Note off	G5	79	48	02:01:048	48
1	Note On	E5	76	0	02:01:048	48
1	Note off	E5	76	48	02:02:096	48
1	Note On	C6	84	0	02:03:000	48
1	Note off	C6	84	48	02:03:048	48
1	Note on	G5	79	0	02:03:048	48
1	Note off	G5	79	48	02:03:048	48
1	Note on	E5	76	0	02:03:048	48
1	Note off	E5	76	48	02:03:096	48
1	Note on	D5	74	0	03:01:000	48
1	Note on	D5	74	48	03:01:048	48
1	Note on	F5	74	0	03:01:096	48
1	Note on	F5	74	48	03:01:048	48
1	Note on	A5	74	0	03:02:000	48
1	Note on	A5	74	48	03:02:096	48
1	Note on	F5	74	0	03:02:000	48
1	Note on	F5	74	48	03:03:000	48
1	Note on	D5	74	0	03:03:048	48
1	Note on	D5	74	48	03:03:096	48

APPENDIX C - STRUCTURAL MAPPING of VIDEO to AUDIO ELEMENTS

Aesthetic Element	Video		Audio	
Light	Type	Directional Non-Directional	Rhythm	Staccato Legato
	Mode	High-key Low-key	Key	Major Minor
	Falloff	Fast Slow	Dynamics	Major Minor
Color	Energy	High Low	Dynamics	Loud Soft
	Hue	Warm Cool	Pitch	High low
	Saturation	High Low	Timbre	Brass,Strings Flutes, reeds
	Brightness	High low	Dynamics	Loud soft
Space	Screen size	Large Small	Dynamics Dynamics	Loud (High Energy) soft(low energy)
	Graphic Weight	Heavy(Close up) Light(long shot)	Chord beat	Complex(accented) Simple(unaccented)
	General shape	Regular Irregular	Sound shape (timbre,chords)	consonant Dissonant
	Placement of objects within frames	Labile Stabile	Chords tension	High(dissonant) low(consonant)
	Texture	Heavy Light	Chords	Complex Simple
	Field Density (no of elements in single frame)	High Low	Harmonic density	High low
	Field Density (no of elements in successive frame)	High Low	Melodic density	High low
	Field Complexity in single shot	High Low	Harmonic complexity	High low
	Field Complexity in successive shot	High Low	Melodic contrapuntal density	High low

Space	Graphic vectors	High Low	Melodic Line	Definite Vague
	Index vectors	High Low	Melodic progression	Definite Vague
	Principal vector Horizontal	High Low	Sound vector orientation (Melodic)	Definite Vague
	Principal vector Vertical	High Low	Sound vector orientation (Harmonic)	Complex Simple
Time/motion	Motion vectors Event Rhythm	High Low Even Uneven	Volume, Tempo Sound rhythm	High Low Even Uneven
	Change in field of view(zoom)	Fast Slow	Dynamic	Fast crescendo and diminuendo
	Vector continuity	Good Bad	Melodic progression, Rhythmic continuity	Even Uneven
	Transitions (cuts,dissolves) Rhythm	Abrupt Gradual Complex Simplex	Modulation (change from one key to another) Sound rhythm	Extreme conservative Complex Simple
Aesthetic energy	Vector magnitude	High Low	Dynamics	Loud Soft
	Vector field energy	High Low	Sound vector energy	High Low