Multimedia Analysis and Synthesis

Mohan S Kankanhalli

Departnment of Computer Science
School of Computing
National University of Singapore
69042 Heidelberg, Germany
mohan@comp.nus.edu.sg
http://www.comp.nus.edu.sg/~mohan

Abstract. We describe novel approaches to multimedia analysis and synthesis problems. We first present the experiential sampling technique which has the ability to focus on the analysis task by making use of the contextual information. Sensor samples are used to gather information about the current environment and attention samples are used to represent the current state of attention. In our framework, the task-attended samples are inferred from experiences and maintained by a sampling based dynamical system. The multimedia analysis task can then focus on the attention samples only. Moreover, past experiences and the current environment can be used to adaptively correct and tune the attention model. This experiential sampling based analysis method appears to be a promising technique for general multimedia analysis problems. We then present the multimedia synthesis technique based on analogies. This method aims to synthesize new media objects based on some existing objects on which appropriate transformation can be applied using analogical reasoning. This multimedia synthesis technique appears particularly useful in the area of computational media aesthetics.

1 Introduction

1.1 Multimedia Analysis

Multimedia processing often deals with live spatio-temporal data which have the following attributes:

- They possess a tremendous amount of redundancy
- The data is dynamic with temporal variations
- It does not exist in isolation – it exists in its context with other data. For instance, visual data comes along with audio, music, text, etc.

However, many current multimedia analysis approaches do not fully consider the above attributes which leads to two main drawbacks – inefficiency and lack of adaptability. The inefficiency arises from the inability to filter out the relevant aspects of the

data and thus considerable resources are expended on superfluous computations on redundant data. Hence speed-accuracy tradeoffs cannot properly be exploited. If the ambient experiential context is ignored, the approaches cannot adapt to the changing environment. Thus, the processing cannot adapt itself to the task at hand.

On the other hand, we have solid evidence that humans are superb at dealing with large volumes of disparate data using their sensors. Especially the human visual system is quite successful in understanding the surrounding environment at appropriate accuracy quite efficiently. These attributes in the experiential environments play an important role for the human visual perception to understand the visual scene accurately and quickly. We argue that like in the case of human perception, multimedia analysis should be placed in the context of its experiential environment. It should have the following characteristics: 1. The ability to "focus" (have attention), i.e., to selectively process the data that it observes or gathers based on the context. 2. Experiential exploration of the data. Past analysis should help improve the future data assimilation. In return, these two attributes would help the analysis to deal with the redundancy and diversity of the spatial-temporal data which is particularly important for real time applications.

In order to achieve this, we describe a novel technique called experiential sampling, i.e., sampling multimedia data according to the context. As shown in Figure 1, by sensing the contextual information in the experiential environment, a sampling based dynamic attention model is built to maintain the focus towards the interest of the current analysis task. Only the relevant samples survive for performing of the final task. These samples precisely capture the most important data. What is interesting is the past samples influence future sampling via feedback. This mechanism ensures that the analysis task benefits from past experience.

As an illustrative example of an analysis task, the face detection problem in videos is described. The experiential sampling technique can be utilized for many other applications involving multimedia analysis such as object detection, object recognition, object tracking (face recognition or traffic sign recognition), context aware video streaming and surveillance. For example, in a surveillance application, intruders can enter only from the boundary of the scene. Therefore, when there are no intruders in the scene, the attention of visual analysis should focus on the boundary. If there is an intruder, the focus of attention should evolve to follow the person.

1.2 Multimedia Synthesis

A well-produced video always makes a strong impression on the viewer. However due to the limitations of the camera, the ambient conditions, or the skill of the videographer, sometime the quality of captured videos falls short of expectation, such as those from a war, medical check-up, surveillance or home videos. On the other hand, we have vast amount of superbly-captured videos available on the web and digital libraries. We describe the novel approach of video analogies which provides a powerful ability to improve the quality of videos by utilizing computable video features. We denote the mechanism of video analogies as A:B::A':B'. B is a target video whose quality we wish to improve and A is a high-quality source video having similar con-

tent. During the learning phase, we find the correlation between this pair. We obtain the correspondence by using feature learning. Then for the target video B, we utilize this correspondence to transfer some desired trait of A (which is A') in B in order to synthesize a new video B'. Thus the new video B' will obtain the desired features A' of the source video A while retaining the merits of the target video B. We demonstrate the power of the analogies technique by describing two applications – video rhythm adjustment and audio-video mixing. We describe the details of the technique in each case and provide experimental results to establish the efficacy of the proposed technique.

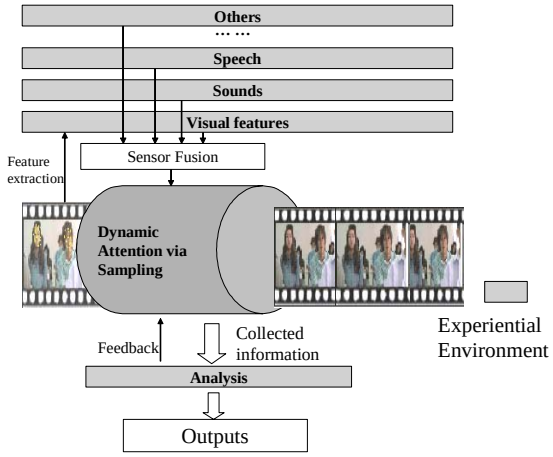


Fig. 1. Experiential sampling for multimedia analysis

2 Experiential Sampling

We would like to stress that the class/style files and the template should not be manipulated and that the guidelines regarding font sizes and format should be adhered to. This is to ensure that the end product is as homogeneous as possible.

2.1 Preliminaries

Experience in multimedia analysis is any information that needs to be specified to characterize the current state of the multimedia system. It includes the current environment, a priori knowledge of the system domain, current goals and the past states. The current goal and prior knowledge provide a top-down approach to analysis. It also determines which features of the visual scene and other accompanying data type should be used to represent the environment. The past states encapsulate the experiences till the current state. More importantly, when we consider the experiential environment, the analysis task needs to systematically integrate the top-down and bottom-

up approaches. In our framework, we allow the analysis to guide the attention on to regions or data of interest from the entire spatio-temporal data.

2.2 Sensor Sampling

Studies on human visual system show that the role of experiences used in top-down visual perception increases in importance and can become indispensable when the viewing conditions deteriorate or when a fast response is desired. In addition, humans get information about the objects of interest from different sources of different modalities. Therefore, when we analyze one particular data type (say spatio-temporal visual data) in multimedia, we cannot constrain our analysis to this data type only. Sensing other accompanying data like audio, speech, music, and text can help us find out where is the important data. Therefore, it is imperative to develop a sampling framework which can sense and fuse all environmental context data for the purpose of multimedia analysis. The current environment is first sensed by uniform random sensor samples and based on context, we compute the attention samples to discard the irrelevant data. Spatially, higher attended samples will be given more weight and temporally, attention is controlled by the total number of attention samples.
In the sampling framework, we represent the environment et at time *t* as:

$$e_t = \{S(t), A(t)\} \tag{1}$$

The environment et comprises of sensor samples $S(t)$ and the attention samples $A(t)$. The sensor samples are basically uniform random samples at any time *t* which constantly sense the environment. The attention samples are the dynamically changing samples which essentially represent the data of interest at time *t*. The attention samples are actually derived dynamically and adaptively at each time instance from the sensor samples in our framework through sensor fusion and the assimilation of the past experience. Once we have the attention samples, the multimedia analysis task at hand can work only with these samples instead of the entire multimedia data. These focused attended samples are the most relevant data for that purpose. The cues for obtaining the context in the environments can be classified as temporal cues and spatial cues. They can be visual features extracted from the current processing visual data or information from their accompanying data (speech, sounds, text etc.). Basically, sensors can sense these cues in order to infer the state of the environment.
In our framework, $S(t)$ is a set of $N_S(t)$ sensor samples at time *t* which estimates the state of the multimedia environment. These sensor samples are randomly and uniformly generated. Since we do not change the number of the sensor samples with time, we will drop the time parameter and N_S denotes the number of sensor samples at any point in time. $S(t)$ is then defined as:

$$S(t) = \left\{s(t); \Pi^S(t)\right\} \tag{2}$$

where s(*t*) depends on the type of multimedia data.

2.3 Attention Sampling

Our sampling based dynamic attention model systematically integrates the top-down and bottom-up approaches to infer attention from the environment based on the context. Thus, the number of attention samples dynamically evolves so the number will be increased when more attention is required and vice-versa. Moreover feedback from the final analysis task is used to tune the attention model with time.

The attention in a scene can be represented by a multi-modal probability density function. Any assumptions about the form of this distribution would be limiting. However, not making any assumption about this distribution leads to intractability of computation. Therefore, we adopt a sample-based method to represent the visual attention. For example, in the one dimensional case, the visual attention is maintained by N samples $a(t)=[s^1(t),\ldots,s^N(t)]$ and their weights $=[\pi^1(t),\ldots,\pi^N(t)]$ as shown in Figure 2. It provides a flexible representation with minimal assumptions. The number of samples employed can be adjusted to achieve a balance between the accuracy of the approximation and the computation load. Moreover, it is easy to incorporate within a dynamical system which can model the temporal continuity of visual attention.

We represent the dynamically varying $N_A(t)$ number of attention samples $A(t)$ using:

$$A(t)=\left\{a(t);\Pi^A(t)\right\} \tag{3}$$

where $a(t)$ depends on the type of multimedia data. Consider the traffic monitoring application shown in Figure 3, Figure 3 (a) has more motion activity and hence needs more attention samples to represent this motion attention. As shown in Figure 3 (b), 567 attention samples (marked as yellow points) are required to represent this motion attention using our method. In contrast, Figure 3 (c) has less motion and needs fewer attention samples. As shown in Figure 3 (d), no attention samples are needed. We determine the number of attention samples $N_A(t)$ based on the current state and the past experiences.

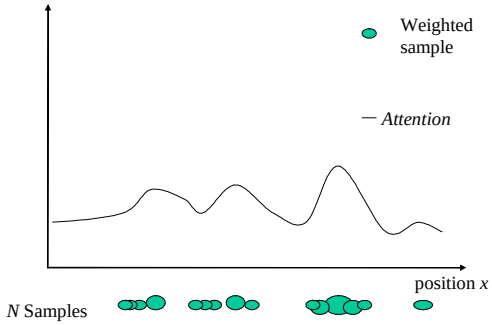


Fig. 2. The multi-modal attention can be represented by N samples $a(t)=[s_1(t),\ldots,s_N(t)]$ and their weights $=[\pi_1(t),\ldots,\pi_N(t)]$.

(a) frame 37 (b)frame 37 (c)frame 479 (d)frame 479

Fig.3. Temporal motion attention.(a) more motion activity (b) 567 attention samples are employed to represent this motion attention.(c) need less attention at this time(d) No attention samples are needed at this time. The number of attention samples is calculated by using equation (8).

Dynamical Evolution of Attention

Attention is inferred from the observed experiences coming from the experiential environments. That is, we try to estimate the probability density of the attention (which is the state variable of the system) at time t using $P(a_t|E_t)$. Note that E_t consists of all the observed experiences until time t which is $E_t=\{e_1,\ldots,e_t\}$, at is the "attention" in the scene and $a(t)$ is the sampled representation of at. Attention has temporal continuity which can be modeled by a first-order Markov process state-space model as shown in Figure 4. The value of at may not be observed though the experience et, which influences the attention at, is observable. In this model, the new state depends only on the immediately preceding state, independent of the earlier history. This still allows quite general dynamics, including stochastic difference equations of arbitrary order. Therefore,

$$P(a_t \mid a_{t-1},\ldots,a_0) = P(a_t \mid a_{t-1}) \tag{4}$$

Based on the above state space model, the a posteriori density $P(a_t|E_t)$ can iteratively be obtained by knowing the observations (likelihood) $P(e_t|a_t)$, the temporal continuity (dynamics) $P(a_t|a_t\text{-}1)$ and the previous state density $P(a_{t\text{-}1}|E_{t\text{-}1})$ as shown in Figure 5.

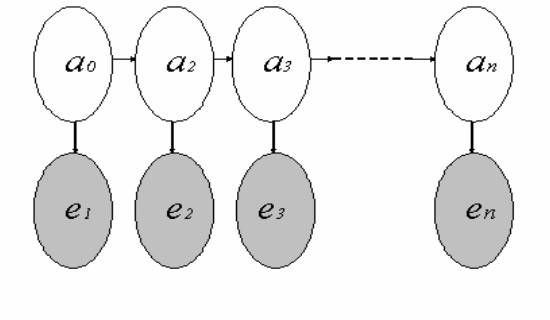


Fig. 4. State-space model for attention.

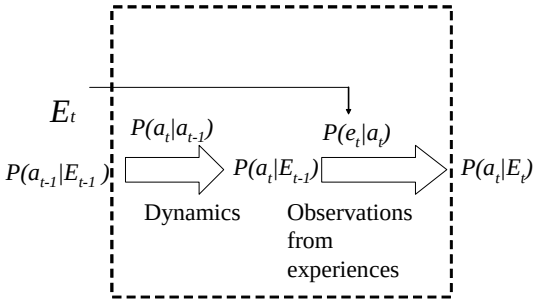


Fig. 5. Iteration of calculating attention state density $P(a_t|E_t)$ in state-space model. By knowing previous state density $P(a_{t-1}|E_{t-1})$ and current experiences et, $P(a_t|E_t)$ can be approximated by a sampling method in the form of samples.

2.4 Experiential sampling technique

In this section, we only provide the outline of the experiential sampling algorithm. The interested readers may consult [3] for details. The ES algorithm can be briefly described as follows:

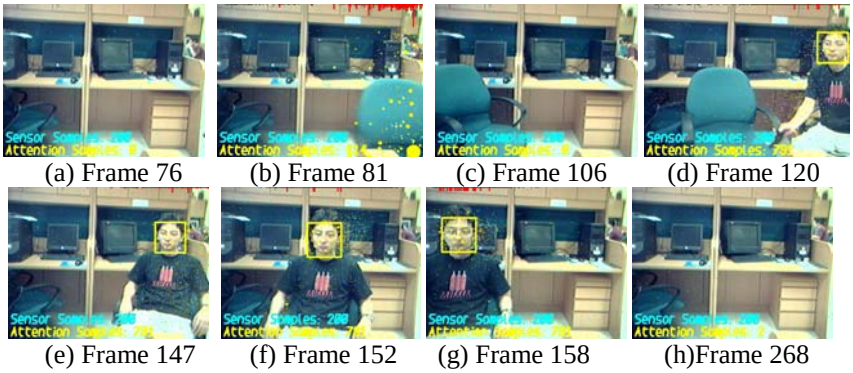
Algorithm: Experiential Sampling (ES):

1. Initialization: t=0
2. {SS(t)} ← Uniform Sampling
3. Asat(t) ← sum of {SS(t)}
4. Ns(t) ←Asat(t)
5. if Ns(t) =< 0; t=t+1; goto step 2
6. {AS(t)} ←Importance Resampling from {*SS*(t)}
7. *for* each *AS*, perform the analysis task
8. t=t+1; goto step 2.

As a general analysis framework, the experiential sampling technique can be used for a variety of multimedia analysis tasks, especially real-time applications like traffic monitoring and surveillance. As a test example, we have applied this framework for the face detection problem in videos. We utilize the context (domain knowledge and accompanying audio (speech) and visual cues (skin color and motion)) to infer the attention samples. These attention samples are adaptively maintained by the sampling based visual attention framework proposed in the previous section. We use the adaboost face detector as the multimedia analysis task. Face detection is only performed on the attention samples to achieve robust real time processing. Most importantly, past face detection results serve as context to adaptively correct the attention samples and the skin color model in the attention inference stage. This allows the face detector to cope with a variety of changing visual environments

2.5 Results

We describe the results of the experiential sampling technique when applied to the face detection problem. Sensor samples are employed to obtain the current visual environment from the skin color, motion and speech cues. The face attention is maintained by the attention samples.



(a) Frame 76 (b) Frame 81 (c) Frame 106 (d) Frame 120

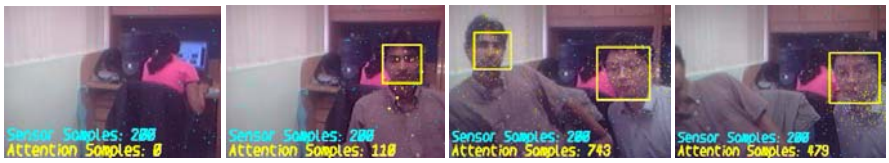
(e) Frame 147 (f) Frame 152 (g) Frame 158 (h)Frame 268

Fig. 6. Face detection sequence 1. (a) static frame N_A=0 (b)A chair moves N_A =414 (c)the chair stopped. N_A =0 (d) a person comes. N_A =791 (e) a person. N_A =791 (f) one person. N_A =791 (g) one person. N_A =791 (h) static frame. N_A =2.

The adaboost face detector is executed only on the attention samples which indicate the most probable face data.

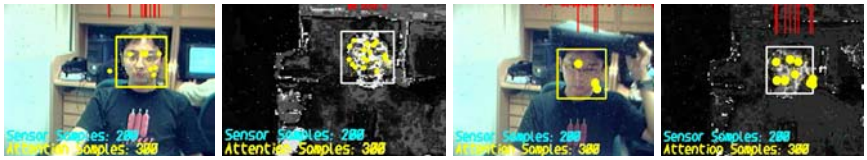
As shown in Figure 6, N_S number of sensor samples is set to 200.The number and spatial distribution of attention samples can dynamically change according to the face attention. In Figure 6(a), there is no motion in the frame, so N_A, the number of attention samples is zero. No face detection is performed. In Figure 6(b), when a chair enters, it alerts the motion sensor and attention is aroused. N_A increases to 414. Face detection is performed on the 414 attention samples. But the face detector verifies that there is no face there. In Fig 6(c) as the chair stops, there is no motion and so the attention samples vanish. In Figure 6(d)-(h) attention samples come on with the face until the face vanishes. Note that depending on how much the attention is, the number of attention samples is different. For instance, N_A in Figure 7 (c) is 743 which is bigger than in Figure 7 (b) and (d) since Figure 7 (c) has two attention areas whereas Figure 7 (b) and (d) only have one. Figure 7 (c) also shows our sampling technique can maintain more than one attention region. Figure 8 (a) is a face under normal light. Figure (b) shows its skin color saliency map calculated by the equation (13). Figure 8 (c) and (d) are a shadowed face and its skin color saliency map. Figure 8 (b) and (d) shows that the feedback of face detector can update the skin color model H_t and make it more adaptive to the visual environment.(note that the skin color saliency map as shown in Figure 8 (b) and (d) is not necessary to be maintained in our method). Figure 9 shows that speech cue can help to create the attention samples when there is not any motion attention initially. As shown in Figure 9 (c), even though speech is off, relevant atten-

tion samples still survive by being given higher weights from the feedback of the previous face detection.



(a) N_A = 0 (b) N_A = 110 (c) N_A = 743 (d) N_A = 479

Fig. 7. Face detection sequence 2.



(a) (b) (c) (d)

Fig. 8. Skin color histogram H_t updated by feedback from the final analysis (face detection).

(a)Frame 2 (b)Frame 10 (c) Frame 42 (d) Frame 100

Fig. 9. Speech experience (a) speech off N_A =0.(b) speech on. N_A becomes 1000. Face detected.(c) speech off. N_A becomes 711(feedback from previous face detection). Face is detected (d) speech on. face detected. N_A becomes 1000..

3 Synthesis Using Analogies

3.1 Preliminaries

We will now describe a method to synthesize multimedia objects using analogical reasoning. The idea is to transfer the desired characteristics from a multimedia object into the given object. We now provide a general description of our scheme with respect to digital video. Given a target video and a designated source video with similar content, we match one common feature of the video pair. Utilizing this feature correspondence, we transfer some other feature of the source video to the target video. For instance, we can build texture correspondence between the video pair and then transfer the color feature of the source video to the target video. We would like to point out

here is that the compared videos should be similar in terms of video content at the shot level so that a proper analogy is set up.

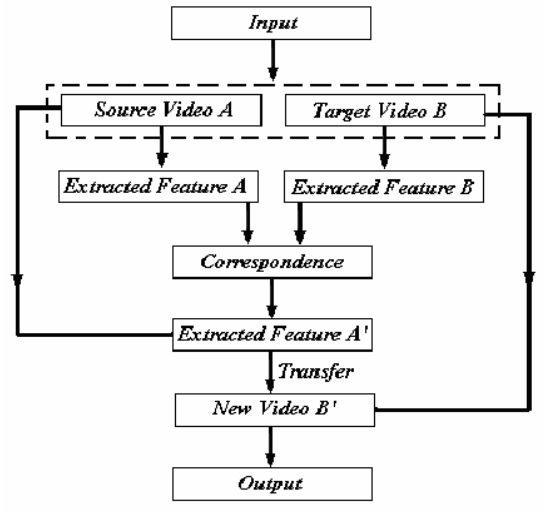


Fig.10. Flowchart for video analogies

The framework of video analogies involves two phases as shown in Figure 10: learning and transfer. The inputs are a source video and a target video. At the learning stage, the corresponding computable features in the source video A and the target video B need to be extracted and compared for similarity (A:B). During the transfer stage, we establish a new function to transfer the desired features (A') from the source video to the target video to create the new video B' by employing the analogy A:B::A':B'. The source video is assumed to be a high-quality clip which we wish to emulate. The target video possibly has some artifacts or shortcomings which we wish to overcome by mimicking the source video. We have observed that our framework of video analogies can potentially be utilized for several problems:

- Color transfer: We can transfer suitable colors from a source video to a target video, such as transferring the source colors to a grayscale (X-ray or infrared) video. This can generate a more vivid and attractive video.
- Texture transfer: By transferring the specific texture to some areas, we can overlap and patch areas on video frame such as annoying video logos or raindrops on glass windows.
- Motion trajectory transfer: By transferring the desired motions to the target video, we can actually remove or reduce shakes caused by camera vibration. Conversely, we also can borrow the shakes from a source video to add excitement to a target video.
- Music matching: Music is an indispensable component of videos. Automatic music matching for atmosphere enhancement is a crucial step in video editing. From the source videos, we can track features in order to add similar music to the target video.

- Aesthetic styles transfer: Artistic styles in a professional movie embody the experiences and knowledge of the directors. Most of these artistic features are at the semantic level; however some of them are apparent at the low level itself such as color, lighting, motion trajectory and rhythm. Such stylistic aspects can be advantageously transferred.

Although video analogies based digital video handling thus can be applied in many situations, here we describe our work on two problems: video rhythm adjustment and audio-video mixing.

3.2 Video rhythm adjustment

Video rhythm refers to the duration and frequency of segmented events, it is subject to the pace of the events and the relationships between these events. The overall rhythm is usually determined by the transitions between video components such as shots, scenes, and sequences. Usual home videos often do not possess a proper aesthetic rhythm. In order to emphasis special scenes, actions, characters and atmosphere in a video, video rhythm adjustment is extremely important. In this section, we discuss video rhythm adjustment based on video analogies. Our motivation is that video rhythm adjustment can be used to emphasize the content and significantly enhance the atmosphere.

In order to perform video rhythm adjustment, we work at the video clip level. Given a source video V_1 and a target video V_2, we segment them into several clips based on the events which are subject to our needs and obtain the clip lengths (total frame number). Suppose V_1 has n clips: S_1, S_2,…, S_n, V_2 has m clips: s_1, s_2,…, s_m, we can find the clip proportion of V1(S_1: S_2:…: S_n). From this ratio, we can determine the relative duration or frame numbers of the video clips. We use it to modify the target video by keeping the ratio invariant. Namely:

$$C_i = C_{i-1} \cdot S_i/S_{i-1}, (m>i> 2) \quad (5)$$

where S_i is the new length of video clips, S_0 can be fixed based on the requirements, for instance, $C_0= S_0$.

In order to automatically detect the rhythm changes in a video shot, we take the video motion into account. We subtract two adjacent frames in the shot. i.e. $\Delta_i = |F_j-F_{j-1}|, F_j \in V_i$, $j<S_i$. If the motion in a clip is a lot, the difference will be significant, or else the distinction is minor. Thus we can find these minor differences by calculating the density of minor difference in this portion. i.e. $\Omega(\Delta_j) = \sum_{T>\Delta i} / \sum_{j<Si}$. If a density is the highest one among all segmentations of this clip, we think the rhythm in this portion is slow.

In practice, we need to add frames or drop frames from the clips of target video according to equation (6):

$$d_i = C_i - s_i \quad (6)$$

If d_i >0, then we need add di frames in the i-th clip. New frames can be created by frame interpolation or replication. If d_i <0, then we need drop di frames from the i-th clip. We add or drop video frames according to equation (7).

$$C_{ij} = [j * C_i / s_i], j =1,2,\ldots, s_i \quad (7)$$

where function [•] is the floor function.

The procedure to drop frames from a clip is rather easy, however the procedure to add frames for a clip is difficult, requiring interpolation and frame synthesis. In our implementation, we just replicate the adjacent frames to expand the sequence which is sufficient in many cases. Our algorithm can now be described as follows:

Algorithm (Video rhythm adjustment):

1: Input a source video and a target video;

2: Segment the source video into several clips according to its events and rhythm;

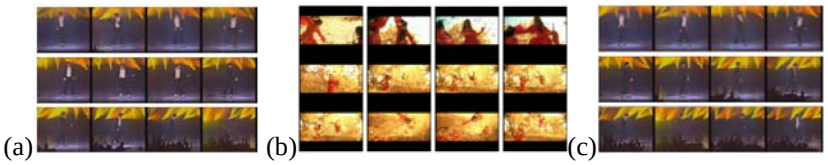
3: For automatic detection of rhythm in a video clip, calculate the density of minor difference of video frames. Determine the length ratio of the two clips;

4: Calculate the new ratios of the target video by using video analogies according to (5);

5: Calculate the new frames of the target video using (7).

6: Output the new synthesized video.

Figure 11 depicts two groups for video rhythm adjustment. Given a video clip from a movie (a), we map the proportion of clip length to the target video (b), we get an exciting video clip (c). Group 1 is a video about Michael Jackson's dance performance to which we transfer the rhythm of a piece of clip of the movie "Hero". This transfer emphasizes the dancing skills of Michael Jackson. Group 2 is the video about creaming of Bill Gates, to which we transfer the rhythm of one clip of the movie of Jackie Chan (Flying Motorcycle) and obtain a new vivid clip that accentuates the creaming effect.



Group1

Group2

Fig.11. Video rhythm transfer

3.2 Video rhythm adjustment

There have been efforts in the recent years to make home videos look more pleasing to viewers by mixing it with appropriate music. Most of the existing software enables the user to add music of his preference. It assumes that the user has enough knowledge about the aesthetic mixing principles. We use the analogies technique to add audio to video by synthesizing appropriate music based on the video content. We

have developed a system that takes in music examples selected by the user and generates new music by applying the aesthetic rules of audio-video mapping. We have A (an example) and B (the output pitch profile based on the given video). We need to transfer certain aspects of A (melody, rhythm, etc.) to B. For transferring melodies, we derive an underlying A' and pose the problem as A':A::B:B'. Given the hue profile of the video, we arrive at the melody profile for video that closely corresponds to the example melody track. The derived melody profile follows the contours of the example.

The music examples in our experiments are melodies mainly selected from western classical instrumental music. Figure 12 gives the pitch contour of a melody. The profile in solid line indicates the original pitch and the profile in dotted line is the Haar wavelet approximation of the music example. The pitch contour of the video as shown in Figure 13 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. The matched contour is shown in Figure 14. A user survey done on the analogy results suggests that the music generated is acceptable though some parts of music may seem repetitive or may not be musically pleasing. This analogies based method in general is preferred over the rule based generation of chord music with which we had experimented earlier.

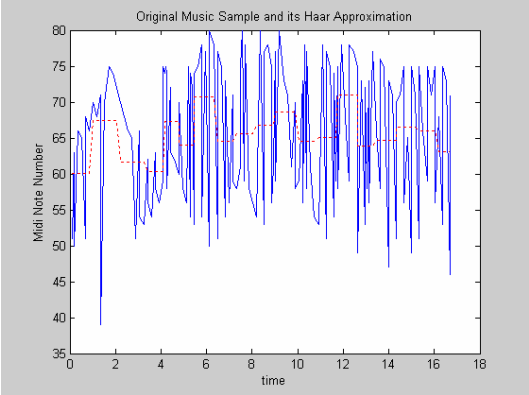


Fig.12. Pitch Contour of a Bach melody

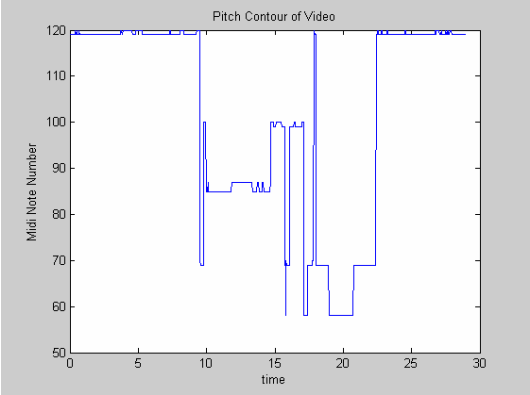


Fig.13. Pitch Contour of 'airplane' video clip obtained from sonification layer.

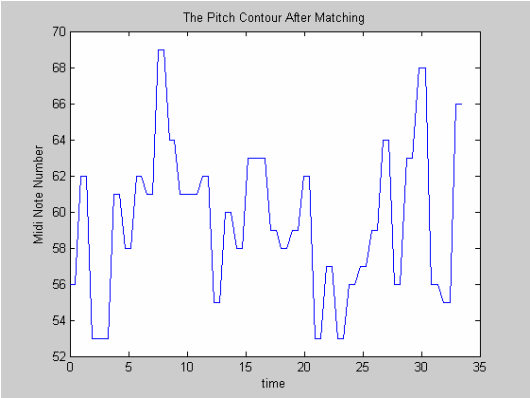


Fig.13. Pitch Contour of synthesized music.

4 Conclusion

We have described a novel sampling based framework for multimedia analysis called experiential sampling. Based on this framework, we can utilize the context of the experiential environment for efficient and adaptive computations. This technique can be extended for general multimedia processing when operating with multiple data streams with possibly missing data. Moreover, it can be incorporated into a dynamical feedback control system for continuous systems. The analogy synthesis technique is a powerful multimedia synthesis technique which can be used for many consumer entertainment applications. It is particularly useful in the new area of computational media aesthetics which seeks to establish the computational foundations of media aesthetics. This is a particularly exciting interdisciplinary area of research.

5 Acknowledgement

The work described in the paper has been done in collaboration with international colleagues: Jun Wang (Delft University of Technology), Wei-Qi Yan (National University of Singapore), Ramesh Jain (GeorgiaTech), S H Srinivasan (Satyam Computer Services), Meera Nayak (National University of Singapore) and Marcel Reinders (Delft University of Technology). Their help and contributions have been invaluable.

References

1. Nayak M., Srinivasan S.H., and Kankanhalli M.S., Music Synthesis for Home Videos: An Analogy Based Approach, Proc. IEEE Pacific-Rim Conference on Multimedia (PCM 2003), Singapore, December 2003.
2. Mulhem P., Kankanhalli M.S., Hassan H., and Yi J., Pivot Vector Space Approach for Audio-Video Mixing, IEEE Multimedia, Vol. 10, No. 2,(2003) 28-40.
3. Wang J. and Kankanhalli M.S., Experience Based Sampling Technique for Multimedia Analysis, Proc. ACM Multimedia Conference 2003, Berkeley, November 2003.
4. Wang J., Kankanhalli M.S., Yan W.Q., and Jain R.. Experiential Sampling for Video Surveillance, Proc. First ACM Workshop on Video Surveillance 2003, Berkeley, November 2003.
5. Wang J., Yan W.Q., Kankanhalli M.S., Jain R., and Reinders M.J.T., Adaptive Monitoring for Video Surveillance, Proc. IEEE Pacific-Rim Conference On Multimedia (PCM 2003), Singapore, December 2003.
6. Yan W.Q., Wang J., and Kankanhalli M.S., Video Analogies. [manuscript under preparation]