AUTOMATIC GENERATION OF DUBBING VIDEO SLIDES FOR MOBILE WIRELESS ENVIRONMENT

Wei Wang and Michael R. Lyu
Dept. of Computer Science & Engineering
The Chinese University of Hong Kong
Email: {weiwang@nudt.edu.cn, lyu@cse.cuhk.edu.hk }

ABSTRACT

Mobile wireless video delivery is still challenging due to its limited bandwidth and dynamic channel status. In this paper, a novel approach named Dubbing Video Slides (DVS) is proposed to cope with the bandwidth limitation problem. Based on a statistical video content importance analysis, DVS method can dynamically select and transmit representative video frames which are relatively more important, and discard others according to current network status feedback. To save bandwidth, we can use these representative frames as substitutes for those adjacent video intervals and synchronize with the original audio track during playback. The visual simulation shows DVS works well for video summary in mobile network delivery.

1. INTRODUCTION

With the rapid development of mobile wireless network and popularization of wireless terminals, versatile mobile service requirements also increased rapidly. Among them, video delivery is the most important one. As a major objective pursued by the communications manufacturers, video delivery via mobile wireless network faces diverse challenges[1][2], including limited bandwidth, dynamic network conditions with low stability, variety of relay equipment, different terminal decoding speeds, various display screen resolution and color depth, confliction between high power consumption, and limited battery capacity, etc. Research work aiming at these challenges has been conducted recently and some achievements have been made including more efficient transmission protocols across different network layers, open interface standards to the Internet backbone, more efficient data compression encoding and decoding, better transmission control, improved error correction, adaptive QoS control, and efficient power control, etc.

Even with all these improvements, video delivery based on MPEG4 and RTSP cannot satisfy the actual requirement yet. Restricted wireless bandwidth is a dominating factor. Compromise must be made when providing video service at present. It would be better to discard those unimportant frames selectively rather than dropping frames passively and randomly during delivery. Therefore, video summary [3][4] becomes an attractive approach for the current mobile wireless video services. Existing solutions mainly focus on video skimming and static video storyboard [5], which were designed to help rapid browsing for locating what users want in a large video database. Static storyboard just provides visual outline without providing the audio information. Video skimming is composed of most brilliant clips without involving the whole video. Given specific videos, both of them produce fixed summaries. In our context, however, a new scheme is needed which should both include audio-visual information and reflect the outline of the whole video. Furthermore, it should be able to produce summaries in different granularity according to the dynamic variety of network bandwidth. This sets forth to the design and implementation of our video summary scheme.

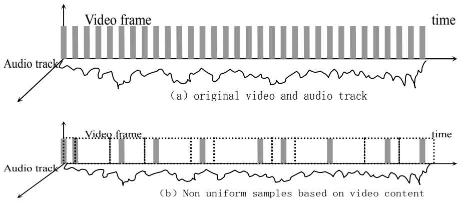
Video frame
time
Audio track
(a) original video and audio track
Video frame
time
Audio track
(b) Non uniform samples based on video content

Figure1: Original Video and DVS

2. Dubbing Video Slides (DVS) Scheme

We describe an innovative summary scheme named Dubbing Video Slides (DVS) for mobile wireless video delivery, which can leverage dynamically between bandwidth and video quality. As shown in Figure 1, via DVS we can select and transmit dynamic amounts of representative video frames which are deemed more important and discard others, based on a statistical video content importance analysis. We then use these representative frames as substitutes for those adjacent video intervals, and synchronize them with the original audio track during playback. As long as the discarding of unimportant frames is restricted to a local scope, and the synchronization with audio track is conducted according to corresponding positions in the sampling sequence, users can still comprehend the delivered content by means of prior knowledge and local context.

The key problem then is to select frames dynamically which are the most representative to the whole original video sequence when synchronized with audio track. As shown in Figure 2, Our DVS generation consists of four steps: (1) The whole video is segmented into basic clip units. (2) Content feature vectors are extracted and the frame sequence is translated into a high dimension trajectory composed of the feature vector points. (3) Trajectory characteristic is analyzed, and more predictable points are discarded based on a dynamic network parameter. A simplified trajectory of more important representative frames is obtained, which can visually represent the outline of the original video. (4) The selected frames are transferred and synthesized with the original audio track during playback. We mainly describe the preceding three steps in the following section in detail.

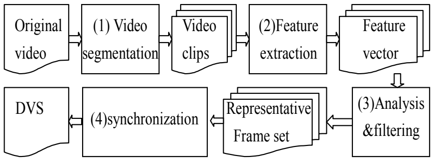

Figure 2: Steps of Automatic DVS Generation

3. DVS GENERATION STEPS

3.1. Video Segmentation

In the first step, whole video is segmented into basic semantic clips. Extracting representative frames from each local clip scope can guarantee that at least one representative frame is selected for each local scene so that no major scenario is missing, which accordingly can achieve the comprehensiveness effect of DVS. Two methods are combined to segment a video into clips.

First, an enhanced double-threshold shot detection is executed which can suppress the over-segmentation of shots as well as reduce the probability of missing shots effectively [6]. Caption-based segmentation is then executed according to whether captions exist or change inside the shots. Video captions, especially dialogue captions, are synthesized manually and they have strong semantic synchronization relationship with the audio track; therefore, they are good clews for DVS summary. Representative frames of different caption videos should still keep this semantic relationship. Caption-based segmentation utilizes text detection techniques such as edge or corner detection, effect enhancement, projection, etc. [7] to estimate whether a video frame contains text-like captions area and locate their positions. In the DVS context, captions usually appear in a rectangle area at the bottom of the screen with overlapping mode, and a certain aspect ratio of the fonts. With this knowledge in advance, we can judge whether a video frame contains text captions and then partition the shots accordingly. As for clips containing captions, text area sub-image is extracted from the whole frame one by one via procedures such as gray image transform, noise filtering, binary image transform, and single character segmentation, OCR is then applied to the normalized single character binary image and the complete character string can be recognized.

Our experiment shows that such a text detection technique is precise enough to locate captions in video frames, but the OCR results are not satisfactory due to complex background and low image resolution. Our original successful OCR rate is about 50% which is not good enough for video content index, but it is good enough to distinguish frame sequences with different captions. As a result, we obtain finer granularity than that of shots which contain either the same caption or no caption at all.

3.2. Fuzzy Color Histogram Feature Vector

To carry out content-feature-based selection of the representative frames, appropriate structural feature vector for each frame which can represent or distinguish video content is formed to analyze the frame similarity relationship in the feature space. As an application-independent method, color-based features, especially color histograms are widely used to construct such vectors with reasonable computation complexity. But there are clear objections in basic color histogram due to rigid color region partitions and sparse pixel statistics. Since human visual and mental perception of color difference is based on continuous shift and not sensitive to approximative colors, and human perception of images is determined by the percentage and distribution of less dominant colors, an improved fuzzy color histogram algorithm based on fuzzy classification is proposed to extract the fuzzy feature vector which better matches human visual perception. This algorithm is now briefly described.

Assuming that a given frame is a color image with width W and height H. We show how to construct V, the corresponding color feature vector. Partition the three independent color channels into n^h, n^s, n^v intervals respectively, and obtain the following partitioned results:

$$C^h = \left\{C^h{}_1, C^h{}_2, ..., C^h{}_i, ..., C^h{}_{n^h}\right\}$$
$$C^s = \left\{C^s{}_1, C^s{}_2, ..., C^s{}_j, ..., C^s{}_{n^s}\right\}$$
$$C^v = \left\{C^v{}_1, C^v{}_2, ..., C^v{}_k, ..., C^v{}_{n^v}\right\}$$

For a given pixel P in the frame, assume its value $x = (x^h, x^s, x^v)$ where x^h, x^s, x^v represent color values in the three channels respectively. As an example, we show

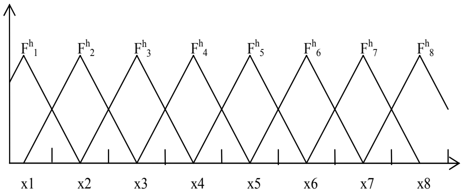

Figure 3: Membership Degree Function

how x^h is classified fuzzily into n^h intervals of C^h. Apply a membership degree function $F^h{}_i(x)$ for each class $C^h{}_i$ $1 \le i \le n^h$. For the given x^h, calculate the results of $F^h{}_i(x^h)$ for every $C^h{}_i$ respectively, and then obtain the n^h dimension membership degree vector, whose items mean to what degree each x^h is subject to the corresponding class, as follows:

$$c^{x^h} = \left\{c^{x^h}{}_1, c^{x^h}{}_2, ..., c^{x^h}{}_i, ..., c^{x^h}{}_{n^h}\right\}$$

As to the membership degree function, there are many forms possible. Triangular function adopted in this paper is shown below:

$$F^h{}_i(x) = \begin{cases} 1 - \dfrac{|x - x_i|}{\varphi} & , \quad |x - x_i| < \varphi \\ 0 & , \quad |x - x_i| < \varphi \end{cases} \tag{1}$$

The graph of Eq(1) is shown in Figure 3. Note that x_i is the value of the middle position in the ith interval, the output of the function is limited to the interval [0,1], and φ means half of the width of the triangular hemline. We can employ such functions to the three individual channels in a similar way:

$$F^h{}_i\left(x^h\right); \quad F^s{}_j\left(x^s\right); \quad F^v{}_k\left(x^v\right)$$

Based on these functions, we can obtain the class membership degree vectors of all pixels in a frame, and then figure out the following fuzzy color histograms for individual channels statistically:

$$H^h = \left(H^h{}_1, H^h{}_2, ..., H^h{}_i, ..., H^h{}_{n^h}\right)$$
$$H^s = \left(H^s{}_1, H^s{}_2, ..., H^s{}_j, ..., H^s{}_{n^s}\right)$$
$$H^v = \left(H^v{}_1, H^v{}_2, ..., H^v{}_k, ..., H^v{}_{n^v}\right)$$

where

$$H^h_i = \frac{1}{W * H} \sum_{\forall x \in G} F^h{}_i\left(x^h\right) \quad , \quad i = 1,2,..., n^h \tag{2}$$

$$H^s_j = \frac{1}{W * H} \sum_{\forall x \in G} F^s{}_j\left(x^s\right) \quad , \quad j = 1,2,..., n^s \tag{3}$$

$$H^v_k = \frac{1}{W * H} \sum_{\forall x \in G} F^v{}_k\left(x^v\right) \quad , \quad k = 1,2,..., n^v \tag{4}$$

The whole color space is partitioned by C^h, C^s, C^v into n subspaces, denoted as:

$$C = \{C_1, C_2, .., C_l, ..C_n\}$$

where $l = 1,2,..,n$, and $n = n^h * n^s * n^v$

The algebraic relation among l and i, j, k can be formulated as follows:

$$l = n^i * n^j * (k - 1) + n^i * (j - 1) + i \tag{5}$$

According to fuzzy mathematics multiplication, we can formulize the fuzzy color histogram feature vector of the given frame as follows:

$$V = \left(H_1, H_2, ..., H_l, ..., H_n\right), \quad n = n^h * n^s * n^v$$

where $H_l = H^h{}_i * H^s{}_j * H^v{}_k$ and

$$1 \le l \le n, \; 1 \le i \le n^h, \; 1 \le j \le n^s, \; 1 \le k \le n^v \tag{6}$$

3.3. Computation of Frame Numbers

Before selection of important frames, the number of extracted frames for each semantic clip should be determined. Assume that the total number of extracted frames of the whole video is N_α which varies dynamically according to the real time feedback of wireless network status, and the total number of clips is ρ.

For a given semantic clip $D^\varphi_{i,j}$, $1 \le \varphi \le \rho$, its frame sequence begins from the ith frame of the original video to the jth frame. Obviously, the number of important frames for a given clip is not only related to the clip length, but also influenced by the intensity of content motions of the clip. Content intensity can be estimated based on the corresponding feature vectors from $V(i)$ to $V(j)$ and represented by I^ψ, where

$$I^\psi = \frac{1}{j - i + 1} \sum_{x=i+1}^{j} \left\| V(x) - V(x - 1) \right\| \tag{7}$$

Using length and content intensity influence factors, we can experimentally define the following extraction ratio of a given clip $D^\varphi_{i,j}$:

$$\beta^\psi = (j - i + 1)^{\frac{1}{2}} * I^\psi \tag{8}$$

The corresponding relative frame extraction ratio is obtained by :

$$\beta^\psi{}' = \frac{\beta^\psi}{\sum_{k=1}^{\rho} \beta^k} \tag{9}$$

Finally, the number of representative frames for $D^\varphi_{i,j}$ can be determined:

$$N^\psi = \left\lfloor N_a * \beta^\psi{}' + 1 \right\rfloor \qquad 1 \le \psi \le \rho \tag{10}$$

3.4. Frame Selection Based on Trajectory Analysis

A reasonable solution for frame selection is to delete those secondary frames which are similar to the previous neighboring ones and whose contents can be estimated from the local context, while keeping those primary frames which have more important visual clews and whose contents are relatively more difficult to foresee. The corresponding feature vector of a given frame can be thought as a point in a high dimension space. Further more, the whole sequence can be mapped into a trajectory composed of these points. This trajectory is composed of connected line segments between discrete and adjacent points, and its shape reflects the content motions of the video. Intuitively, positions with higher curvature in the trajectory refer to the more important frames whose contents are more difficult to deduct, while positions with lower curvature refer to the less important frames whose contents are easier to deduct from the local context. After eliminating those secondary points, the sequence of the remaining points can still present the approximate profile of the original video well.

Assume that the number of representative frames varies in the following range according to the network status.

$$N_{\min}{}^{\varphi} \leq N^{\psi} \leq N_{\max}{}^{\varphi} \qquad (11)$$

For a given point and feature vector $V(k)$, we define a local context relativity measure $LR(V(k))$, which equals to:

$$\left|\left\|V(k+1)-V(k-1)\right\|-\left(\left\|V(k)-V(k-1)\right\|+\left\|V(k+1)-V(k)\right\|\right)\right| \qquad (12)$$

Its value reflects the degree of predictability of a certain point. It is intuitive that the above three points form a triangle in a hyper plane. $LR(V(k))=0$ means that point k is on the line between points $k-1$ and $k+1$, and the content variety at point k is not intense, so it can be easily inferred from the local context. Otherwise, if point k departs farther from the line between points $k-1$ and $k+1$, then $LR(V(k))$ will be greater, which means the local variety is greater. After $LR(V(k))$ of all the points are figured out, we sort them and delete the points with the minimal values.

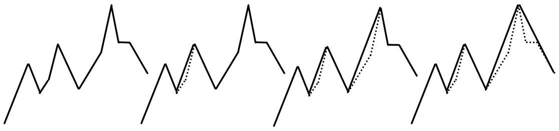

Figure 4: Sketch Map of Trajectory Simplification

As shown in Figure 4, we repeat the above operation on the sequence of the remaining points, until the number of points left equals the given minimal value $N_{\min}{}^{\varphi}$, and then complete the frame selection process.

4. EXPERIMENTS

Based on these steps, we calculate a content importance weight for each video frame. By comparing those weight values with a dynamic threshold whose value reflects the current network status, we can generate dynamically sequences with different lengths from several tens of frames to the total length of the video. We performed several experiments to investigate different granularity of DVS summaries whose selected video frame number decrease gradually, thus reducing the required bandwidth by sacrificing the visual continuity. Due to the length limitation, location data of these representative frames in the testing videos are ignored here.

We use Adobe Premier to simulate the final visual effect of the resulting DVS instead of implementing the transmission and synchronization. Such process is subjective and hard to evaluate, but from simulation experiments, we can still find that although the image conversion result is not fluent and visual details are lost, the integrated audio-visual outline of the original video is comprehensible and valuable. With our testing material, even after 96% frames are discarded, viewers can still grasp the outline of the video. In addition, we find that with an appropriate frame number, DVS can also generate quite similar key frame storyboard as mentioned in [5].

5. CONCLUSION

To cope with the bandwidth problem in mobile wireless video delivery, this paper proposed a dynamic frame selection approach for dubbing video slides creation. It can select dynamically representative video frames which are relatively more important, and discard others according to current network status feedback. Visual simulation experiments show that DVS indeed can provide a simple and feasible video content creation and playback technique for quality video delivery under current mobile wireless communications constraints.

6. ACKNOWLEDGEMENT

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK4360/02E).

7. REFERENCES

[1] S. Gamze, "Challenges of Wireless Media Streaming", *Proc. International Conference of SSGRR*, L'Aquila, Aug 2001.

[2] J. Vass, and S. Zhuan, "Scalable, Error Resilient, and High-Performance Video Communications in Mobile Wireless Environments", *IEEE Trans. Circuits and Systems for Video Technology*, July 2001, pp. 833-847.

[3] E. Minoru and S. Shun'ichi， "MPEG-7 enabled Digest Video Streaming over 3G Mobile Network", *Proc. International Conference on Packet Video*, Nantes, France, 2003.

[4] B.L. Tseng, C.Y. Lin, and J. R. Smith, “Video summarization and personalization for pervasive mobile devices”, *Proc. International Conference on Storage and Retrieval for Media Databases*, SPIE, 2002.

[5] Ying Li, T. Zhang and D.Tretter, "An overview of video abstraction techniques", *HP Laboratory Technical Report*, HPL-2001-(191), July 2001.

[6] Rainer Lienhart, "Reliable Transition Detection In Videos: A Survey and Practitioner's Guide", *International Journal of Image and Graphics (IJIG)*, March 2001, pp. 469-486.

[7] M. Cai, J.Q. Song and M.R. Lyu, "A New Approach for Video Text Detection," *Proc. International Conference On Image Processing*, New York, USA, 2002.