VIDEO BOOKLET

Xian-Sheng HUA, Shipeng LI, Hong-Jiang ZHANG

Microsoft Research Asia
{xshua; spli; hjzhang}@microsoft.com

ABSTRACT

In this paper, we propose a novel system, *Video Booklet*, which enables efficient and nature video browsing and searching. In the system, a set of selected thumbnails excerpted from the original videos are printed out on a real physical booklet or album. When users plan to browse the content of or search a specific clip in their digital video library, they can firstly browse their booklets in such a manner as browsing a typical papery family album. When he wants to watch the segment represented by a certain thumbnail in the booklet, he is able to use his camera phone to capture the corresponding thumbnail. Then the captured image is sent to computer or other devices connected to the monitor via wireless network, and last the Video Booklet system will automatically find the corresponding segment in the video library for the user and begin to play the segment. Video Booklet builds a seamless bridge between digital media library and analog papery albums.

1. INTRODUCTION

The quantity of multimedia data, in particular, home videos, is increasing dramatically in recent years with the popularity of digital camcorders. Unlike text data, which is much easier to be indexed and randomly accessed, it is well-known that this is not any easy task for media data. It is time-consuming, as well as much more inconvenient, for users to search, retrieve and browse their personal media data. Though there are quite a few media content indexing and browsing systems available in literature, most of them are mainly working on video segmentation, scene grouping, summarization, or specific event detection And the user interfaces of most of these systems are based on PCs, which means, the users are required to use keyboard, mouse or remote control to locate or search the content they want to see in the storage, according to the navigation hints on the computer screen or TV monitor.

In this paper, we propose a novel system, *Video Booklet*, which enables nature and efficient video library browsing and searching. As illustrated by Figure 1, the system consists of two sub-systems: Video Booklet Generation and Booklet-Based Video Browsing and Searching. In the first sub-system, videos are segmented into scenes, shots and sub-shots, as well as the signature and a set of features are extracted from these segments. Then based on these features, the temporal structure, and user selected booklet template, a video booklet is generated and printed out using a typical LaserJet color or black and white printer.

In the second sub-system, when users want to browse the content of their digital video library, they can firstly browse their booklets in a manner as browsing an ordinary family album. When they wants to watch the segment indicated by a certain thumbnail in the booklet, they can use their camera phone, or web cam, or other typical and convenient capture devices to capture the corresponding thumbnail in a leisure manner. Then the captured image is sent to the computer or other device that is connected to the monitor via wireless network, and the Video Booklet system will automatically find the corresponding segment in the video library and begin to play the segment. The system builds a bridge towards seamless communication between digital media world and analog papery world.

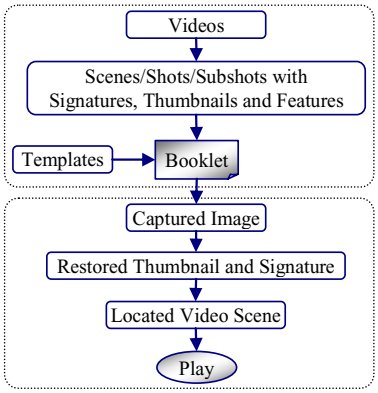


Figure 1. Flow Chart of Video Booklet.

The key difficulties in Video Booklet lie in: (1) How to generate the booklet, including how to select an appropriate set of thumbnails and how to organize the thumbnails in the printout that facilitates efficient browsing and search. (2) How to match the captured thumbnail with the media clip in the video library in the case of that the captured images are often in poor quality and greatly distorted.

A most related research work to Video Booklet is proposed in [2], in which a video paper system is presented. It enables news video and meeting video browsing by scanning a bar code on the printouts. The main difference between this work with Video Booklet is that we use image signatures instead of printing and scanning bar codes, thus the user interface is more nature for home video applications. Furthermore, in Video Booklet, an optimization-based thumbnail selection method is proposed, as well as booklet templates are supported, which enables generating impressive, personalized, and flexible printouts. Another related commercial work is linking pictures on magazines, newspapers to web advertisements by embedding and extracting watermarks [1]. However, extra information is needed to be embedded in the pictures, as well as typically high-quality print/press is required, thus not suited for general home users.

The rest of the paper is organized as follows. Section 2 introduces preliminary video content analysis algorithms. Booklet generation is detailed in Section 3, followed by video browsing and searching with booklet in Section 4, and experiments in Section 5. Conclusions are presented in Section 6.

0-7803-9332-5/05/$20.00 ©2005 IEEE

2. VIDEO CONTENT ANALYSIS

Before generating and using Video Booklet, content analysis is applied on the entire video library, which consists of three primary steps, video structuring, feature extraction, and signature extraction. For convenience, we describe our system based on one video file, while it is also applicable for all or any subset of the entire video library.

2.1 Structuring and Feature Extraction

The raw home videos are segmented into a three-layer temporal structure, from small to big including subshots, shots and scenes, according to color similarities or timestamp (if it is provided or can be recognized). There are a number of shot boundary detection algorithms in literature [9]. The one we used here is similar to the one in [4] if the videos are analog; otherwise the shot boundaries are directly derived from the discontinuousness of timestamps. Accordingly a video V can be represented as a series of shots, denoted by $V = \{Shot_i, 0 \le i < N_{shot}\}$.

Subshot is a sub segment within a shot, or we may say, each shot can be divided into one or more consecutive subshots. In Video Booklet, subshot segmentation is equivalent to camera motion detection, which means one subshot corresponds one unique camera motion. For example, if in a shot the camera paned from left to right, and zoomed in to a specific object, then paned to the top, and zoomed out, and then stopped, then this shot consists of four subshots including one pan to right, on zoom in, one pan to top, and one zoom out. In our paper, the camera motion detection algorithm in [5] is adopted.

Simultaneously, an "attention" or "importance" value of each subshot is calculated by averaging the "attention indexes" of all video frames in the subshot, in which the attention index is the output of "attention detection" [6] relating to object motion, camera motion, color and speech in the video. Suppose the video is represented by a subshot series as $V = \{Sub_i, 0 \le i < N_{sub}\}$, where N_{sub} is the number of subshots, The "importance" of each subshot is denoted by a_i. And the middle frame of each subshot is chosen as thumbnail (key-frame) of the corresponding subshot. The pictures in the printout booklet are a selected set of thumbnails from these ones (to be detailed in Section 3). The thumbnail of the subshot with the longest duration in the shot is chosen as the thumbnail of the corresponding shot (only used for scene grouping). The quantized color histogram in HSV space [6] of the thumbnail is taken as the similarity feature of the subshot and shot. In addition, a visual quality measure of Sub_i, denoted by q_i, is derived from a set of visual features (contrast, gray and color histogram), which is similar to the quality measure in [4].

Then the shots are grouped into scenes, while scene segmentation is constrained by users' inputs. Actually Video Booklet is a scalable representation of users' video library. Each scene is represented by one thumbnail in the printout album, and scene segmentation can be coarse or fine, depending on users' choices. To be exact, the parameters that users may choose include desired number (or range of number) of scenes, average duration of the scenes, and average dissimilarity between different scenes. Users may choose any of these parameters to constraint scene segmentation, as detailed below.

Suppose the video collection that is selected to be printed out will be segmented into K scenes according to content similarity and timestamp (if available). In particular, we define the similarity of any two consecutive shots (or we may say it is the similarity measure of the "connection point" of these two shots) as the weighted sum of histogram intersections of the shots at the two sides of the "connection point":

$$Sim_i = Sim(Shot_i, Shot_{i+1}) = \sum_{j=1}^{\delta} \beta_j S_{i,i+j} + \sum_{j=-\delta}^{-1} \beta_{-j} S_{i+j+1,i+1} \quad (1)$$

where $0 \le i < N_{shot}$-1. $S_{k,l}$ is the histogram intersection (a color similarity measure) of $Shot_k$ and $Shot_l$ in HSV space, and the parameter β_j is the summing weight defined by

$$\beta_j = \beta^j, 0 < \beta \le \delta \quad (2)$$

in this implementation ($\beta = 2/3$, $\delta = 5$).

Scene segmentation is equivalent to finding a set of "cut points" in series $\{Sim_i\}$. If the scene segmentation is constrained by the number of scenes (K), assume we will "cut" the shot list at "connection point" list θ, while θ is an K-element subset of $\{0, \ldots, N\text{-}2\}$, and Θ the set of all subsets of this form. "Connection points" in Sim_i whose subscripts are in θ are the "cut points". Then, we grouped the shots into K groups by solve the below optimization problem:

$$\theta^* = \arg\min_{\theta \in \Theta} \frac{\sum_{j \in \theta} Sim_j}{|\theta|} \quad (3)$$

where $|\theta|$ stands for the number of elements in the finite set θ. *Genetic Algorithm* (GA) [8] is employed to approach global optimal solution for this optimization problem. If scene segmentation is constrained by the average scene duration, or the dissimilarity between scenes, a similar method is applied. The main difference is the feasible solution set. Consequently, a video V can also be represented by $V = \{Scene_i, 0 \le i < N_{scene}\}$.

2.2. Signature Extraction

The ordinal measure of the thumbnail is extracted and regarded as the signature of each subshot, similar to that of [3]. Ordinal measure reflects the relative intensity distribution within an image, which was first proposed in [7] as a robust feature in image correspondence. In our system, the key-frame (thumbnail) of each subshot is partitioned into $N = N_x \times N_y$ blocks and the average gray level in each block is computed. Then the set of average intensities is sorted in ascending order and the rank is assigned to each block (in this paper, N_x=N_y=5). The $N_x \times N_y$ dimensional rank sequence, i.e., ordinal measure, is the inherent relative intensity distribution in the frame, thus is naturally robust to color distortion caused by print and scan. Furthermore, ordinal measure is a very compact feature vector ($25 \times 5/8 < 16$ bytes/subshot if we use 5 bits to represent number from 0 to 24).

Consequently, a subshot Sub_i is represented by a feature vector

$$Sub_i = (st_i, et_i, q_i, a_i, s_i) \quad (4)$$

where (st_i, et_i) is the start and end timestamp, q_i and a_i are visual quality and attention measure, respectively. And s_i is the signature of the key-frame.

3. BOOKLET GENERATION

This section describes how to generate a booklet for a collection of video data based on video content analysis that has been introduced in previous section.

3.1 Thumbnail Selection

In this step, one representative thumbnail for each scene is selected from the subshot thumbnails in the corresponding scene. This selected thumbnail is the one that will be printed out on the

booklet (album). The objectives of thumbnail selection are threefold. One is to maximize the average signature differences of all pairs of selected thumbnails, thus the retrieval error could be reduced. The second is to maximize the visual quality of the selected thumbnails, and the third is to maximize the representativeness of the thumbnails, i.e, maximize the average attention index. This issue is formulated as follows.

Suppose the subshot in $Scene_i$ is represented by $\{Sub_{i,j}=\{ st_{i,j}, et_{i,j}, q_{i,j}, a_{i,j}, s_{i,j}\}, 0 \le j < N_i\}$. And suppose the corresponding subshot of the selected thumbnail of this scene is denoted by $Sub_{i,j(i)}$, where $0 \le j(i) < N_i$. It is obvious that there are $\prod N_i$ possible feasible solutions (denoted by Θ). Suppose one of the feasible solutions is denoted by θ. Then the average signature difference of selected thumbnail is

$$SD_\theta = \frac{1}{C^2_{N_{scene}}} \sum_{0 \le m < n < N_{scene}} Dist\left(s_{m,j(m)}, s_{n,j(n)}\right) \quad (5)$$

where $Dist(\cdot,\cdot)$ is the distance of two signatures defined in [3]. Similarly, the average visual quality and representativeness are represented by equation (6) and (7), respectively.

$$VQ_\theta = \frac{1}{N_{scene}} \sum_{0 \le i < N_{scene}} q_{j(i)} \quad (6)$$

$$RP_\theta = \frac{1}{N_{scene}} \sum_{0 \le i < N_{scene}} a_{j(i)} \quad (7)$$

Therefore, the thumbnail selection problem can be formulated as an optimization problem:

$$\max F(\theta) = \alpha SD_\theta + \beta VQ_\theta + \gamma RP_\theta \quad (8)$$

where $\alpha, \beta, \gamma \ge 0$, $\alpha + \beta + \gamma = 1$. This is a mathematical programming problem. The problem is easy to be rewritten as a 0-1 programming problem, and thus is easy to be solved by GA or similar heuristic searching algorithms.

To obtain better results, several constraints may be added to problem (8), such as the minimum of signature difference (SD_θ).

3.2 Booklet Template

To provide impressive printout booklets in a variety of forms, a series of booklet templates are designed. With booklet templates, for the same selection of thumbnails we can generate different forms of booklets. Booklet templates only affect the layout and look-and-feel of the thumbnails in the booklets in current system since complex templates may require complex thumbnail detection algorithm in the step of browsing and searching, as to be explained in Section 4. Figure 2 shows three simple booklet templates samples.

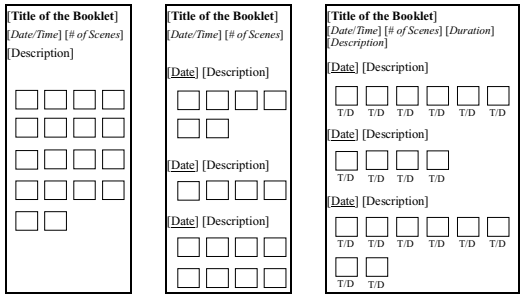


Figure 2. Video Booklet Templates (T/D: Time/Duration).

4. BROWSING WITH VIDEO BOOKLET

In this section, we will show how to accomplish nature and efficient video browsing/searching through the booklet generated in previous section. The primary step is to locate the actual image boundary of the captured thumbnail. Thereafter, the detected image area is restored to its original shape by perspective transformation, and then the signature is extracted. At last the most similar thumbnail in the video library will be retrieved and the system will play the video from the beginning of the scene that the thumbnail represents. Figure 3 shows some typical samples of leisurely (without much effort on adjusting the focus area when doing capturing) captured thumbnails.

Figure 3. Samples of Typical Captured Video Thumbnails.

As mentioned above, the first step is to locate the boundary of the target thumbnail. To accomplish this goal, the following steps are applied (the first captured image in Figure 4 is taken as an example).

(1) *Edge Detection*: After enhancing the contrast, the edge map of the captured image is detected using Canny edge detector, and then converted into 0-1 image where the pixel value is 1 if it is edge; otherwise is 0 (Figure 4(a)).

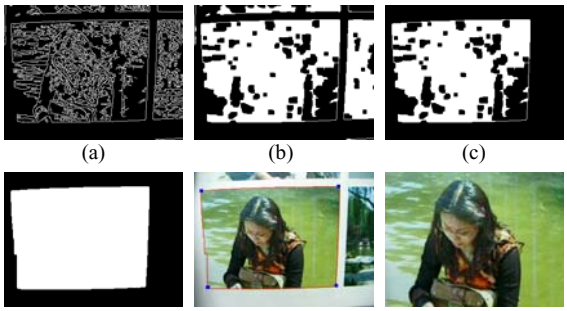
(2) *Dilation and Erosion*: Three rounds of dilation and erosion (mathematic morphology operations) are applied on the map, thus the close nonzero pixels are merged together to form blocks (Figure 4(b)).

(3) *CC*: After that, for each nonzero pixel in the *centric area* (a rectangle in the center that covers half area of the entire map), finding its connected component (CC) and convert all zero pixels within its CC into nonzero. Therefore, more pixels in the target frame area are converted to nonzero pixels. For original nonzero pixels that do not connected to any of the detected CC are convert to zero pixel, thus the noises near the boundaries are eliminated (Figure 4(c)).

(4) *Flood Fill*: Then to make nonzero area covers more portion of the target thumbnail, the zero pixels who at least has 3 nonzero neighbors (not necessary to be adjacent) in the four directions (left, top, bottom, and right), are converted into nonzero pixels (Figure 4(d)).

(5) *Finding Contour*: The convex contour of the nonzero area is extracted. This contour is the rough edge of the target thumbnail in the captured image. And the four vertices of the contour are extracted by finding dominant corners (Figure 4(e)).

(6) *Refinement*: Sometimes, the contour extracted in previous step has obvious errors, such as one or more corners might be missed, as illustrated by Figure 5(a), which is the result after applying previous five steps on the second image in Figure 3. In this case, the vertices of the *minimum bounding quadrangle* of the nonzero area are applied. To be exact, if the detected vertex is relatively far from the corresponding vertex of the bounding quadrangle, it is replaced (Figure 5(b)).

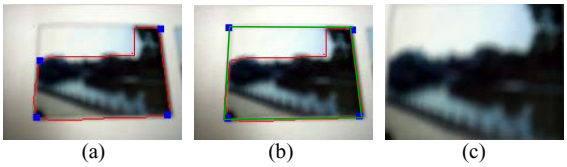


(a) (b) (c)

(d) (e) (f)

Figure 4. Medium Results of Detecting Target Thumbnail.

After the four vertices are detected, the target area is converted into a rectangle (width-height ratio will not affect the signature extraction) using perspective transformation. Figure 4(f) and 5(c) show the results after applying this step. After that, the signature is extracted and compared with the ones in the library, and the closest one in the database will be retrieved. The system then plays the corresponding scene for the user.



(a) (b) (c)

Figure 5. Vertex Refinement.

5. EXPERIMENTS

We test our system on ten-hour home videos of various genres (birthday, Christmas, vacation, wedding and so on). There are totally 4876 shots and 13082 subshots. Ten booklets are generated from different portion of the video library. For each printed thumbnail in each booklet, 100 queries are applied to retrieve the corresponding scene in the library. Among the 100 query samples, 20 are captured by a typical digital camera in its lowest resolution (640×480), 40 of them are from a low-end webcam, and the rest from a typical camera phone (0.3M pixels).

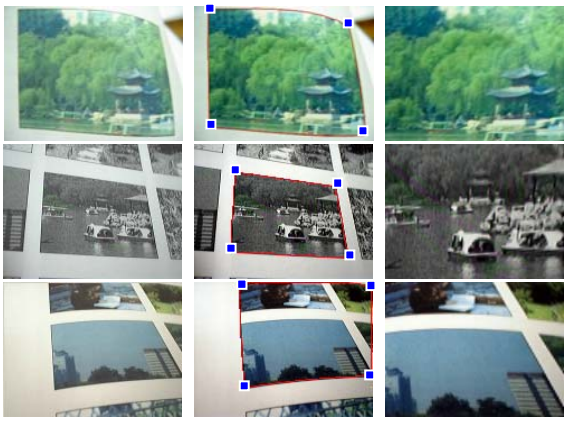


Figure 6. More Samples (*Left: captured images; Middle: detected contour; Right: Restored images***).**

Table 1 illustrates the retrieval accuracy, which shows Video Booklet system has very high retrieval accuracy even for the low-end webcam (which creates poor-quality images). The row labeled with *Acc** is the accuracy without applying thumbnail selection scheme (the thumbnail of the longest subshot is selected as the thumbnail for the scene), which shows the scheme significantly improved the performance.

Table 1. Retrieval Accuracy.

ID	1	2	3	4	5	6	7	8	9	10	Avg
Scene #	84	124	96	84	100	160	200	196	200	300	154.4
Acc*	0.91	0.92	0.99	0.83	0.91	0.92	0.91	0.85	0.82	0.83	0.89
Acc	0.98	0.99	1.00	1.00	0.98	0.98	0.95	0.95	0.97	0.93	0.97

Note: *ID* - Booklet ID; *Scene #* - Number of thumbnails; *Acc*: Accuracy; *Acc** - Results without using thumbnail selection scheme.

Samples in Figure 3 are all corrected retrieved. Figure 6 shows two more succeeded and one failed sample. The failed ones mostly are due to the contours of the target frames are not correctly located.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel system for browsing and searching personal video clips, Video Booklet, which builds a seamless bridge between digital video library and analog papery albums. There are a number of extensions for Video Booklet. First, similarity search might be applied according to gray or color histograms. Thus users are able to find similar clips in the video library. Second, the system is also applicable for photo library to enable quick and convenient access of the photo collections. In this case, one photo is considered as one subshot or shot. Third, after correctly located the corresponding scene, we may play a short summary of the scene, thus AVE [4] can be adopted into the system. Fourth, if the distinguish capacity of the signature is not sufficient, a hierarchical booklet could be applied, in which the user may capture a booklet ID image first, and then the signature comparison could be restricted into the specific range of video collection in the database.

Future work is to implement the above extensions, as well as design more robust frame detection algorithm and image signature to deal with the cases when the templates are more complex. For example, when the background is graphics or picture, and/or thumbnails are in round or ellipse shape or any other shapes, or even they are overlapped partially. These kinds of booklets are more impressive while much more difficult to be corrected retrieved.

7. REFERENCES

[1] Digimarc, http://www.digimarc.com/.

[2] J. Graham, *et al*, "The Video Paper Multimedia Playback System," *ACM Multimedia 2003*.

[3] X.-S. Hua, *et al*, "Robust video signature based on ordinal measure," *IEEE Intl. Conf. on Image Processing 2004*.

[4] X.-S. Hua, L. Lu, and H.-J. Zhang, "AVE – Automated Home Video Editing," *ACM Multimedia 2004*.

[5] D.-J. Lan, *et al*, "A Novel Motion-Based Representation for Video Mining," *Intl. Conf. on Multimedia and Expo 2003*.

[6] Y.-F. Ma, *et al*, "User Attention Model based Video Summarization," *to appear in IEEE Trans. on Multimedia*.

[7] R. Mohan, "Video sequence matching," *Intl. Conf. on Audio, Speech and Signal Processing 1998*.

[8] D. Whitley, "A Genetic Algorithm Tutorial," *Statistics and Computing*, Vol. 4, 64-85, 1994.

[9] TREC Video, http://www-nlpir.nist.gov/projects/trecvid/.