An Interactive SpiralTape Video Summarization

Yong-Jin Liu, *Senior Member, IEEE,* Cuixia Ma, Guozhen Zhao, Xiaolan Fu, *Member, IEEE,*
Hongan Wang, Guozhong Dai, Lexing Xie *Senior Member, IEEE*

***Abstract*—A majority of video summarization systems use linear representations, such as rectangular storyboards and timelines at linear scales. In this paper, we propose a novel nonlinear, dynamic representation called SpiralTape that summarizes a video in a smooth spiral pattern. SpiralTape provides an unusual and fresh activity suitable for stimulating environments such as science and technology museums, in which children or young individuals can have enjoyable experiences that create meaningful learning outcomes. In addition, SpiralTape provides an uninterrupted overall structure of video content and takes design principles including compactness, continuity, efficient overview and interactivity into consideration. A working SpiralTape system was developed and deployed in pilot applications and exhibition. Elaborate user studies with evaluation benchmarks on multiple metrics were conducted to compare SpiralTape with two representative linear video summarization methods and a state-of-the-art radial video visualization. The evaluation results demonstrate the effectiveness and natural interaction performance of SpiralTape.**

***Index Terms*—Video summarization, user interaction, video content analysis, user experience.**

I. Introduction

RAPID advances in the technologies of pervasive multimedia capture devices,massive storage and network distribution have led to the rapid growth of video data resources. Video summarization aims to present the contents of videos in a concise form so that users can grasp the essence of a long video clip in a short time. In our study, we emphasize that interaction is important in the design of video summarization, because video summarization methods intended to provide an efficient and customized tool for communicating messages among people.

In terms of the users' interactions with video data, video viewing and browsing can be distinguished as two separate phases. The aim of video viewing is to display video content frame-by-frame on a 2D screen. In contrast, in addition to viewing, video browsing gives users interactive controls for viewed video content, e.g., play, pause, fast forward, playback, seek, skip-to-begin and skip-to-end. However, these traditional browsing operations are not efficient because at any time only a single frame is displayed on the screen. Therefore, a new form of interaction should be considered in the design of video summarization to support efficient and customized browsing.

Many new techniques have been proposed to improve video browsing efficiency. Based on a similarity-based interactive search, a video can be viewed as a cloud of distance-driven images [1]. Browsing among the images in a cloud representation can offer users a high degree of freedom that they can use to optimize their tasks and actions. Other structured forms such as interactive and continuous temporal zoom [2], motion-based dynamic narratives [3], scene structure graph [4] and sketch graph [5], have also been proposed to support efficient video browsing operations. All these browsing techniques are designed to fulfill one or a few specific tasks, such as video retrieval, video authoring or motion event analysis. The more general purpose of browsing to understand video content has not been fully considered in these techniques.

Videos are a typical form of time series data. In daily life, people become accustomed to linear reading order from left to right and from top to bottom. Accordingly, linear patterns such as rectangular layouts (e.g., traditional mosaics or storyboards) and timelines at linear scales (i.e., where a unit of distance is equal to a set amount of time) are currently dominant and will likely continue to be in many multimedia applications of video summarization. In our study, we focus on unusual and inventive representations that are often desirable in stimulating environments to provide novel and memorable activities. The application scenarios include exploration (e.g., science and technology museums, scientific content such as survey outtakes), entertainment (e.g., visual puzzle hunt) and novel visualizations for online and smart-TVs, etc. In these scenarios, curious and energetic children or young individuals have enjoyable experiences that create meaningful learning outcomes. Motivated by this notion of enjoyable multimedia learning, in this paper we propose a novel and unusual spiral pattern to summarize video content. We choose the spiral pattern not only because of its aesthetic appeal and compact layout, but also because it supports an efficient overview and an easy-to-use, natural multi-touch interactions. Further details about design principles are presented in Section III.

Although a spiral order has been successfully used in massive time-series data visualization due to its organic appearance [6], [7], [8], [9], to the best of the authors' knowledge, a nonlinear representation using a spiral pattern has not been applied to video summarization. Because video summarization methods aim to assist people to quickly understand video content and fulfill particular tasks via browsing, their interactions with people are important. Therefore, in addition to an unusual and novel nonlinear representation, user's enjoyable interactive experiences are important to spark children or young people's

This work was partially funded by Royal Society-Newton Advanced Fellowship, The Natural Science Foundation of China (61322206, 61521002, 61232013, U1435220) and TNList Cross-discipline Foundation.

[1]State Key Lab of Computer Science, Institute of Software, Chinese Academy of Sciences, China.

[2]Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China.

[3]State Key Lab of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, China

[4]Research School of Computer Science, The Australian National University, Australia

[5]Beijing Key Lab of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, China

SpiralTape Rectangular tiles Tapestries [2]

Fig. 1. Three video summarization patterns with black background are used in an elaborate user study. More details are presented in Section VII and in supplemental demo video.

interest and motivation. The *SpiralTape* proposed in this paper (Figure 1 left) provides such a novel form of video summarization that uses a hierarchical and smooth spiral representation to summarize video content.

We make the following contributions in this paper:

- SpiralTape utilizes a spiral rotation pattern that has sufficient continuity to represent an unbroken timeline structure in a video. We present a novel method to construct SpiralTape from both static and dynamic/interactive perspectives. Compared to traditional linear patterns, the new video-browsing mechanism in SpiralTape can provide a visually pleasing, continuous and uninterrupted overall structure of large-scale video contents.
- The user interface provided in SpiralTape allows users to personalize their video browsing naturally and intuitively using familiar gestures. Accordingly, SpiralTape's gesture-based interactions can help users better understand the video contents in personalized ways.
- A working system including interaction was constructed and deployed in pilot applications and exhibitions. Evaluation benchmarks on multiple metrics were collected. These benchmarks show that the developed system outperforms two representative linear summarizations and one state-of-the-art radial visualization.

II. Related Work

Video summarization is an important tool designed for providing concise representations of video data. It typically generates a condensed summary of a long video clip using either a sequence of still images (e.g., keyframes) or some types of moving images (e.g., video skimming) that allow users to browse videos efficiently and obtain a basic understanding of the content in a brief period. Many state-of-the art video summarization methods provide novel ways to view and browse videos. Broadly there are two classes [10]: (1) static summarizations using keyframes or storyboards and (2) dynamic summarizations using video skimming, synopsis, continuous temporal zooming or motion mosaics.

A number of static summarization methods have been proposed. The majority of these methods synthesize a composite still image from a collection of selected images. For example, Rother et al. [11] synthesized a digital tapestry from a large collection of different images by solving a multi-class labelling problem, and Mei et al. [12] presented a compact synthesized collage for a video sequence in a few different layouts. In addition to photo-realistic still image representation, sketch-like schematic storyboard [4], [13], [14] and comic-book-like pictorial layouts [15] have also been proposed for static video summarization. However, these layouts did not consider the appropriate hierarchical structures for representation and operation when the video is long and contains a large number of keyframes.

Dynamic video summarization has received increasing attentions recently. Video skimming [16] extracted salient information from video contents and reorganized it into a shorter skim video. Video retargeting methods [17], [18], [19] can also be used to generate a shorter skimming video by regarding time as one dimension. Video synopsis [20] is another short video representation that preserves essential activities in the original video. Although these summarizations included dynamic information such as motion events, users can view them only passively, i.e., they cannot interact with the summarizations to actively find the contents that a specific user might want to view.

Most aforementioned summarization methods utilize linear or rectangular layouts. In this paper, we study a spiral pattern that is intrinsically aesthetically pleasing. Although spiral representations have not been applied in video summarization, we pay attention to information visualization, in which concentric, spiral and Euler are known collectively as radial space filling (RSF) patterns [21], [22]. These patterns arrange the data into concentric circles, tight spirals or evenly spaced clusters, and thus utilize space efficiently [22]. In particular, Carlis and Konstan [23] and Weber et al. [7] suggested that spirals are suitable for visualizing time series data with periodic structures. Furthermore, Carlis and Konstan [23] showed that spiral visualizations make it easier to detect certain patterns in the data than traditional line-plot visualizations. A recent work [24] applied a radial layout to visualize the hierarchical structure in a video. We compare this radial video visualization with our spiral pattern in Section VII.

Both video summarization and radial visualization have considered interactivity. Liu et al. [5] proposed a sketch graph that uses gesture operations to quickly and naturally explore user intention in organizing video contents. Herranz et al. [25] proposed comic-like summaries with a context scalability representation. Barnes et al. [2] proposed video tapestries, which

summarized a video using a multi-scale image with continuous temporal zooming. In video tapestries, users can interactively zoom to any desired position to explore the fine-scale temporal details. All these summarizations applied traditionally square- or line-plot diagrams. Interactivity also plays an important role in radial visualization. Most radial visualization methods [22] have supported one or more of five operations proposed in [21], including selection, reconfiguration, distortion, drill-down/roll-up, and pan/zoom/rotation. However, radial layouts have seldom been explored in video visualization.

III. SpiralTape Design Principles

SpiralTape is a visually pleasing, concise video summarization that conveys message in video contents to people. SpiralTape focuses on aiding users to easily understand video contents in a short time and having a good user experience. In the section, we present design principles that help distinguish the visual representation in SpiralTape from traditionally linear patterns.

Compactness. Spirality is an attractive and efficient pattern for visual representation. The Archimedean spiral (also known as the arithmetic spiral) possesses a smooth aesthetic trait that is also compact and has a space-saving characteristic. In a screen display with limited size, SpiralTape can present more video contents (Figure 1a) than traditional line-plot patterns, such as video tapestries [2] (Figure 1c). However, compactness does not entail maximizing the filled area. For example, we can break the line of video tapestries at the screen border and tile the keyframes in multiple lines, as shown in Figure 1b. We call this *a rectangular-tiled pattern*. Obviously, rectangular tiles maximize the filled area on a screen. However, due to inevitably multiple brokens at the borders, it has poor continuity, which is the next principle.

Continuity. Compared to the rectangular-tiled pattern that is inevitably broken at the end of each row, the SpiralTape pattern has greater continuity because the spiral curve is C^{∞} smooth, indicating that the timeline is continuous without any interruptions. Furthermore, the continuity of on-screen exploration combined with visually smooth transition in a hierarchical structure (Figure 4) enhances user experience during interaction.

Efficient overview. SpiralTape presents an overall view of video contents due to its compactness. In general, users prefer to quickly grasp the entire content structure on one screen. Unlike previous representation patterns that typically use a left-to-right and top-to-bottom reading order, SpiralTape supports a new viewing concept in which reading order progresses from the spiral center to the spiral perimeter along different angles.

Interactivity. Different users may have different intentions when attempting to understand video content. User interaction can help users to better understand video summarization in personalized ways; therefore, the design of SpiralTape includes a gesture-based interface. Integrated with the spiral pattern, SpiralTape provides simple and intuitive gestures that support natural interactions, helping users actively browse the video content efficiently.

IV. Static Spiral Construction

An important preprocessing step in SpiralTape is to select keyframes from video clips and organize them into a hierarchical video structure (Section IV-A). Then, this structure is visualized in an Archimedean spiral layout (Section IV-B).

A. Video Data Preprocessing

To provide an efficient overview and a continuous zooming operation in SpiralTape, we preprocess a video clip into a hierarchy of four levels {events, scenes, shots, keyframes}.

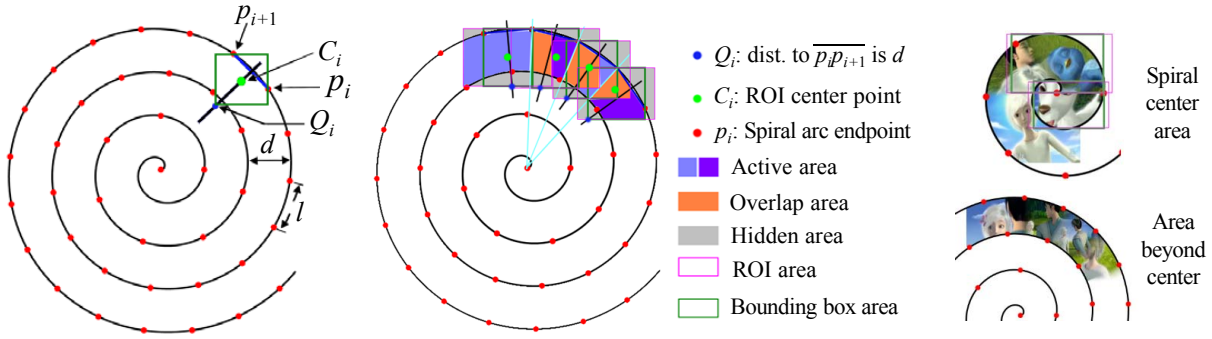
First, we regard the entire video clip as an event. An event consists of a set of scenes, each of which is temporally and spatially cohesive in the physical environment but may not be continuous in video data. Each scene is further decomposed into a set of shots, where a shot is a sequence of consecutive frames that was continuously captured by the same camera [26]. To develop a practical algorithm for building this {events, scenes, shots, keyframes} hierarchy from a video clip, we propose a motion-based frame distance $D(i, j)$ that is inspired by the success of keypoint-based methods [27] for detecting semantic concepts.

Our strategy is that after partitioning the video clips into a set of shots, we cluster the shots into scenes in a high-dimensional feature space. Finally, for each shot, we sample the frames from a uniform time interval and call these frames "keyframes". The motion-based frame distance $D(i, j)$ employs two event characteristics in a video clip [27]: (1) static information in an event that provides what are involved in the event, and (2) dynamic information that provides how an event evolved in the temporal domain.

Denote a video clip as $f(t)$, i.e., a sequence of frames f parameterized by an artificial discrete time t. The SIFT keypoints [28] are computed at each frame and their union over all frames is denoted as Ξ. Each SIFT keypoint in Ξ is a 128-dimensional unit vector v, and the difference between two vectors v_1 and v_2 is measured by 2-norm $\|v_1 - v_2\|_2$. These Ξ can be regarded as a point set in $\mathbb{R}^{128}$. The affinity propagation method [29] is then applied to classify Ξ in $\mathbb{R}^{128}$ into clusters, where the number m of clusters is determined automatically and optimally. The center of each cluster in $\mathbb{R}^{128}$ is treated as a visual word w and the set of points $p \in \Xi$ in this cluster is denoted by $C(w)$. Each point $p \in C(w)$ is an instance of the word w.

Let a video clip be represented by a vocabulary consisting of all its visual words $V = \{w_i\}_{i=1}^{m}$. The vocabulary V characterizes the static information of what are in the video. To characterize the dynamic information of how an event evolved in the temporal domain, the motion information of every instance p of a visual word w is considered. Denote the location of p be (x, y) at frame f_k. The spatial intensity gradient method [30] is applied to track its motion in the next frame f_{k+1}. Denote the resulting motion vector as $m(p)$ which is a 2-vector in image plane. The average motion dispersion vector between two visual words w_i and w_j in V, $i \neq j$, is defined by

$$md(i, j) = \frac{1}{N_i} \frac{1}{N_j} \sum_{r=1}^{N_i} \sum_{s=1}^{N_j} \|m(p_r(i)) - m(p_s(j))\|_2^2$$



(a) Layout determination of one ROI (b) The overlapped layout of four ROIs (c) Layout in and beyond center

Fig. 2. SpiralTape layout scheme. Note that in this scheme, each ROI is always displayed with its original orientation. In other words, the ROI's orientation does not rotate along the spiral curve.

where $p_r(i) \in C(w_i)$ and $p_s(j) \in C(w_j)$, N_i and N_j are the number of instances in the cluster $C(w_i)$ and $C(w_j)$ respectively. The distance between two visual words w_i and w_j is given by

$$wd(i,j) = \lambda \frac{\|w_i - w_j\|_2^2}{\max\{\|w_k - w_l\|_2^2, w_k, w_l \in V\}} + (1-\lambda)\frac{md(i,j)}{\max\{md(k,l), w_k, w_l \in V\}}$$

where λ is a weight balancing the contributions of static and dynamic information of visual words. In our experiment, we choose $\lambda = 0.5$.

If an instance $p \in C(w)$ appears in a frame f_i, we say the visual word w appears in f_i. Denote by $\{w_j(i)\}_{j=1}^{n_i}$ the set of visual words that appear in f_i. We define a semantic distance between two frames f_i and f_j as

$$D(i,j) = \frac{1}{n_i}\sum_{r=1}^{n_i} \min\{wd(w_r(i), w_s(j), 1 \le s \le n_j)\} + \frac{1}{n_j}\sum_{s=1}^{n_j} \min\{wd(w_r(i), w_s(j), 1 \le r \le n_i)\}$$

To decompose a video clip into a set of shots, shot boundary is detected by linearly scanning all the frames. Note that a shot boundary may be either abrupt (hard cuts) or gradual (fades, wipes and dissolves). To detect both abrupt and gradual shot boundaries, we use a sliding window of width $2T$, i.e., the frames in the interval $[i-T, i+T]$ for a frame f_i. Suppose we have detected h shots and the latest shot boundary is detected at the frame f_k (initially $h = 0, k = 0$). To detect the boundary of shot $h+1$, we start at f_{k+T} and move forwards in the direction of increasing time. For each frame $f_{k'}$, $k' \ge k+T$, we compute a salience change $SC(k')$ defined by

$$SC(k') = \max\{D(i,j), i \ne j, i, j \in [k'-T, k'+T]\}$$

If $SC(k')$is larger than a predefined threshold τ, a new shot boundary is detected. In our experiment, we set $T = 20$ and $\tau = SC_{total}/10$, where SC_{total} is the total salience change computed in the whole time interval, i.e., T is the half of the number of frames in the video clip.

After shot boundary detection, a video clip is decomposed into a set of shots $\{s_i\}_{i=1}^{n_s}$. We cluster these shots into scenes. Note that a scene may consist of disjoint shots in the video clip. Denote by $\{w_j(s_i)\}_{j=1}^{n_{s_i}}$ the set of visual words that appear in s_i. Similar to the semantic distance between two frames f_a and f_b, the semantic distance between two shots s_i and s_j is defined as

$$D(s_i, s_j) = \frac{1}{n_{s_i}}\sum_{r=1}^{n_{s_i}} \min\{wd(w_r(s_i), w_s(s_j), 1 \le s \le n_j)\} + \frac{1}{n_{s_j}}\sum_{s=1}^{n_{s_j}} \min\{wd(w_r(s_i), w_s(s_j), 1 \le r \le n_{s_i})\}$$

Given the distance $D(s_i, s_j)$ between any two shots s_i and s_j, we apply the affinity propagation method [29] to classify the shots $\{s_i\}_{i=1}^{n_s}$ into clusters and each cluster represents a scene.

In the hierarchical structure of SpiralTape presented in Section V, a shot s is represented by k keyframes in s, where k is adaptive according to different scales. We reorder the frames in s using an order defined by

$$\Pi = \{\lfloor \frac{jn_s}{i} \rfloor, i = 1, 2, \cdots, n_s - 1, 0 < j < i\} \quad (1)$$

where n_s is the number of total frames in s. If a frame appear twice in this order, the later one is removed, that is, if $\lfloor \frac{j}{i} \rfloor = \lfloor \frac{j'}{i'} \rfloor$, $i' > i$, then the frame $\lfloor \frac{j'}{i'} \rfloor$ is removed. For example, if a shot has 6 frames, $\Pi = \{3, 2, 4, 1, 5, 6\}$. Whenever an adaptive number k of keyframes are needed, the first k frames in Π are used.

Finally for every frame f in the video clip, a region of interest (ROI) is computed with the aid of the salience map $M(f)$ [31], which is defined as the bounding box of the maximal connected component in $M(f)$.

B. SpiralTape Layout

We use the polar coordinate (r, θ) to define the Archimedean spiral by $r = \frac{\theta}{2\pi}d$, where the angle θ is measured in radian and d is the pitch parameter. An Archimedean spiral has the property that any ray emitted from the origin intersects successive turnings of the spiral at points with a constant separation distance d. The SpiralTape is designed as a seamless composition of ROIs using the following steps (Figure 2):

- The spiral curve is divided with a fixed arc length l. Refer to red mark points shown in Figure 2a. When $\theta \in (\frac{5}{2}\pi, 4\pi)$, l is chosen to be $1.5d$. When $\theta \in [4\pi, \infty)$, l is chosen to be d. For the first two arcs started at the spiral center, we empirically set the mark points at $\theta = 2\pi$

and $\theta = \frac{5}{2}\pi$. The mark points separate the spiral into a sequence of arcs $A = \{a_i\}_{i=1}^{n_A}$ of fixed arc lengths. An ROI is mapped to a region determined by an arc $a_i \in A$. The user can set the parameters of screen space size and the pitch to determine how many ROIs of keyframes to be displayed. Denote by n the number of ROIs to be displayed.

- A video clip consists of a set of shots $S = \{s_i\}_{i=1}^{n_s}$. Each shot s_i is weighted by its time duration w_i. Then for each shot s_i, we choose the number of ROIs to be displayed in SpiralTape as $\frac{w_i}{\sum_j w_j} n$.
- ROIs of keyframes are mapped to the arcs A with the order of increasing time. Each arc maps an ROI. Refer to Figure 2a. The Cartesian coordinates (x_i, y_i) and (x_{i+1}, y_{i+1}) of two endpoints p_i and p_{i+1} of an arc a_i are computed by

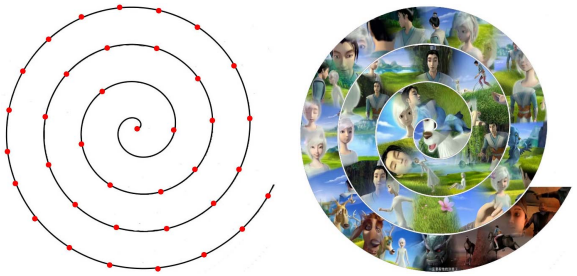
$$p_j = (x_j, y_j) = \left(\frac{\theta_j}{2\pi} d \cos\theta_j, \frac{\theta_j}{2\pi} d \sin\theta_j \right), \ j = i, i+1$$

  By utilizing an auxiliary point Q_i, a rectangular region B_i containing the arc a_i is located as the smallest axis-aligned bounding box of three points (p_i, p_{i+1}, Q_i). The position of Q_i is determined with the following rules. Q_i sits on the bisector of line segement $\overline{p_i p_{i+1}}$ at the side containing the spiral center. The distance from Q_i to $\overline{p_i p_{i+1}}$ is d.
- The n ROIs selected from shots S are displayed in SpiralTape. The ith ROI is mapped to the region B_i by coinciding their centers. Note that ROIs are always displayed with their original orientations. In other words, the ROI's orientation does not rotate along the spiral curve. It can be shown that any two contiguous mapped ROIs always have a small overlapped area (Figure 2b), inside which we blend two ROIs with a gradually changed blending parameter. Therefore, no gaps exist in SpiralTape. Finally, the ROI layout for the first two spiral arcs a_1 and a_2 around the spiral center is treated particularly as shown in Figure 2c. Overall, SpiralTape provides a visually pleasing canvas for video summarization.

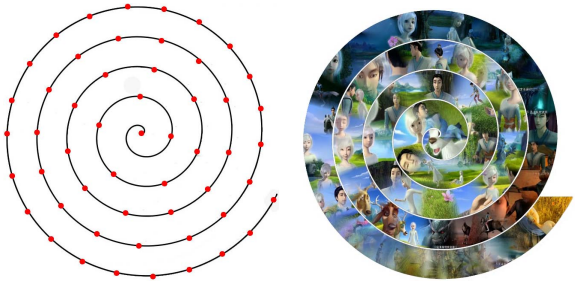
To summarize, the spiral curve is partitioned into arcs of equal length l (which is a function of pitch parameter d) and each arc maps an ROI. In a fixed display area, a smaller value d indicates that more ROIs can be shown and the size of each ROI size is smaller. A user can choose her/his own preferred balance by altering the value of d. See Figure 3 for an example. Similar to Tapestries [2], the hard borders between frames containing ROIs are eliminated in SpiralTape, making it possible to achieve dynamic smooth zooms with spatial continuity (see Section V).

V. SpiralTape Dynamics

SpiralTape is a dynamic video summarization, which supports animation by moving and inserting more or fewer ROIs along the track of the spiral curve. During the animation process, it is time consuming to compute the new position of each moved region B_i in the spiral. To implement a real-time smooth and continuous transition of ROI animation, we sample the angle range of the spiral using a sufficiently small interval and pre-compute the rectangular region associated with each pair of two adjacent sampled angle values.



(a) $d = 4$ units (left) and 40 ROIs mapped (right)



(b) $d = 3$ units (left) and 64 ROIs mapped (right)

Fig. 3. In a fixed display area, a smaller value for the pitch parameter d allows more ROIs to be mapped in SpiralTape, but each ROI is smaller.

In SpiralTape, users can interactively select a spiral arc interval (e.g., in the angle interval (θ_1, θ_2)) to zoom in or zoom out. Taking zooming in for an example, we use the following steps to smoothly move and insert more ROIs into the selected interval $(\theta_1, \theta_2))$ (Figure 4):

- Given the interval (θ_1, θ_2), we first compute the corresponding start point p_1 and end point p_2 in the spiral. During the zooming-in process, p_2 is moved in the direction of increasing angles. Assume p_2 moves to a new position p_2' and there are k ROIs existed between p_1 and p_2, and k' ROIs between p_1 and p_2'. The animation effect consists of two parts: (a) the smooth movement of k already existed ROIs, and (2) the new emergence of $k' - k$ ROIs at a finer scale in the hierarchy specified by Eq. (1).
- The position of each ROI can be indexed by the mark points in the spiral. For each already existed ROI, denote the index of its original position as o_i and the index of its new position after movement as e_i. The number of indices that the ith already existed ROIs should traverse is given by $n_i = e_i - o_i$.
- Different ROIs have different moving velocities. That is, for a $j > i$, the jth ROI should move a bit faster than the ith ROI to provide a smooth effect. To achieve this goal, we define a basic frequency as $N = \min\{n_i\}$, for all i in already existed ROIs. The moving step size of ith ROI is computed by $s_i = \lfloor n_i / N \rfloor$. During the movement of already existed ROIs, if the arc length between any two contiguous ROIs exceeds l, a new ROI is selected using the order defined in Eq. (1) and is inserted into

Fig. 4. Continuous zooming in with hierarchical smooth transition. Red areas represent the parts of smooth transition of ROIs in the animated zooming sequence. Top row: local scaling, in which a piece of spiral arc is zooming in from left to right. Bottom row: global scaling, which shows the overall zooming-in of SpiralTape.

(a) SpiralTape (b) Interactive annotation (c) Hyperlinks (d) Operating on a TabletPC

Fig. 5. SpiralTape provides gesture operations in a user interface. (a) shows a SpiralTape summarization. (b) shows a user's annotation on SpiralTape. (c) shows the gesture-based hyperlink creation. (d) shows a snapshot of operating SpiralTape on a TabletPC. More details are presented in supplemental demo video.

SpiralTape (a red area in Figure 4).

In the demo video submitted with this paper, SpiralTape rotates counter-clockwise when zooming in and clockwise when zooming out. The user can also interactively set customized zooming directions in SpiralTape.

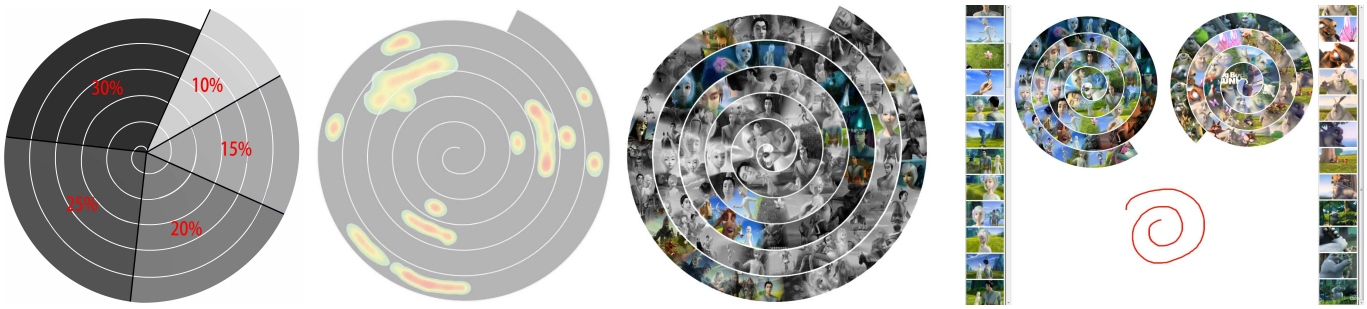
VI. Interaction with SpiralTape

A distinct feature in SpiralTape is to provide natural interactions for users based on gesture operations. A user can interactively operate SpiralTape using simple and intuitive gestures. In addition to the three basic operations, namely, viewing a video by clicking an ROI (Figure 1 left), zooming in and out (Figure 4)), SpiralTape provides four additional operations: annotations, hyperlinks, visual percent of scenes, splitter and joiner. These interactions help users easily understand the video content in a personalized way with the aid of the hierarchical structure {events, scenes, shots, keyframes} established in Section IV-A.

User interface. Both single- and multi-touch gestures are provided in SpiralTape (Figure 5(d)). Using single-touch gestures, a user can trigger a dialogue to create annotations, establish hyperlinks and generate splitter and joiner. Using multi-touch gestures, a user can trigger continuous zooming in or out within the hierarchy of keyframes (Figure 4). By mimicking traditional paper-and-pencil operations, gesture-based user interactions can easily and naturally explore and communicate messages between users and video content.

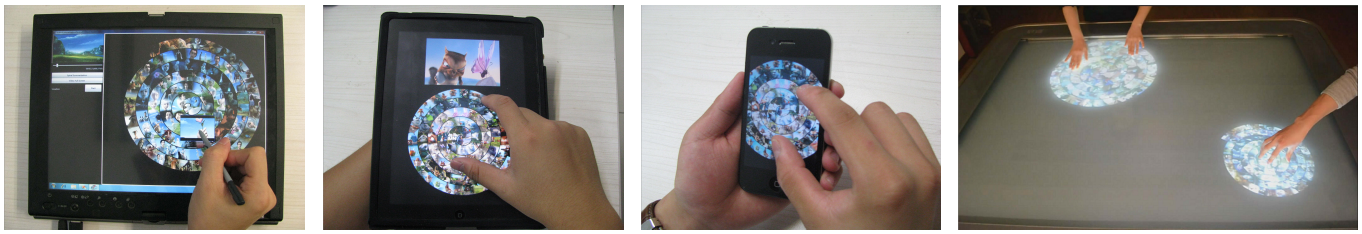
Interactive annotations. A user's interactive annotations on SpiralTape naturally facilitate the creative process by providing personalized annotations during the communication of users. These annotations can be either icons or bookmarks (Figure 5(b)). A user's annotations enrich and enhance video indexing for later video browsing. Subsequently, users can navigate the videos with scalable details by clicking on their annotations. Annotations also reduce the gap between user's high-level semantic concepts and the video's low-level features.

Hyperlinks. A user can create a hyperlink by establishing a relation between two annotations or ROIs. Hyperlinks provide interaction by allowing users to follow these links to navigate among different parts of a video clip or between different video clips (Figure 5c). Compared to traditional timelines that restrict interaction in a linear manner such as play, pause, rewind,



(a) Visual percent of scenes (from left to right: pie charts, heat map, ROI highlight) (b) Splitting and joiner

Fig. 6. (a) SpiralTape supports an intuitive visual presentation of screen time using the area of fan-shapes. Left: the percentages of different scenes in terms of time duration is shown in pie charts. Middle: for any user-specified scene in pie charts, the locations of its shots are highlighted in a heat map, where the heat color is set according to the blending parameters β of ROIs in s, i.e., when β goes to one, the color is hotter and when β goes to zero, the color is colder. Right: by degrading all other ROIs to gray images, those ROIs belonging to the scene are highlight. (b) The operation of splitter and joiner allows users to select shots in one or more videos and reorganize these shots into a new SpiralTape. More details are presented in supplemental demo video.



(a) Tablet PC (b) iPad (c) iPhone (d) Tabletop

Fig. 7. A good scalability of SpiralTape on small mobile devices and large touch-capable devices. More details can be found in supplemental demo video.

fast-forward and step, hyperlink structures can establish a customized relations based on annotated context and video contents in a task-driven manner.

Visual percent of scenes. SpiralTape can take advantage of the spiral pattern to visualize the percentages of different scenes in pie charts (Figure 6(a) left). Pie charts provide users with a quick overview for estimating how many scenes are in a video and the length of each scene. As noted earlier, a scene is temporally and spatially cohesive in a physical environment but may not be temporally continuous in video data. When a user clicks a scene s in a pie chart, SpiralTape shows the locations of its shots in a heat map (Figure 6(a) middle), where the heat color is set according to the blending parameters β of ROIs in s. That is, if $\beta = 1$, the corresponding positions in an ROI of s is completely visible and the color is hot. When β goes to zero, the corresponding positions of an ROI are partially visible (blending with other ROIs) and the color become cold. After users know the locations of shots in a scene, they can switch to an ROI highlight (Figure 6(a) right), which degrades all other ROIs to gray images, leaving only those ROIs belonging to the scene highlighted. Then, a user can better examine the contents in these highlighted ROIs. Visual percent of scenes provides a context+focus representation, leading to an intuitive visual analysis in SpiralTape.

Splitter and joiner. The shots of a scene may be discontinuously located in SpiralTape. A user can first select the shots belonging to one or several closely related scenes using the splitter operation and then reorganize these shots into a new SpiralTape using the joiner operation (Figure 6(b)). Moreover, users can browse the selected shots in detail by zooming in or out using the newly constructed SpiralTape. Together with the visual percent of scenes, the splitter and joiner operation provides an efficient tool for scene selection and understanding content in a task-driven manner.

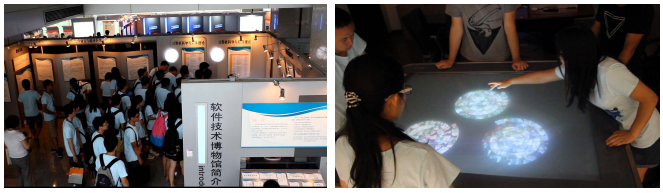
During the training session in a user study presented in Section VII, the combination of interactive annotations, hyperlinks and visual percent of scenes were used to gain an efficient understanding of video content. Furthermore, the combination of visual percent of scenes and splitter and joiner were used for an efficient video content indexing. See Supplementary Material for details.

SpiralTape has good scalability because it is easily extendable to popular small mobile devices, such as iPhone and iPad, and to large touch-capable devices, such as Tabletop. See Figure 7 for examples. On mobile devices with small displays, SpiralTape's compactness and inherently efficient overview lead to a good user experience. On large touchable devices, the hyperlinks and splitter and joiner operations, along with the support of multi-touch gestures, further enhance the user experience.

VII. User Study

We implemented SpiralTape with the proposed user interface in the C# platform and tested it with different display-integrated tablet PCs (Figures 5d and 9) and touch-capable devices such as the iPad, iPhone and Tabletop (Figure 7).

Application scenario. SpiralTape was designed to provide a visually pleasing and novel video summarization that can be used in stimulating environments such as science and technology museums. We exhibited SpiralTape on various



(a) STM at Beijing, China (b) SpiralTape on Tabletop

Fig. 8. Left: the Software Technology Museum (STM) located at Institute of Software, Chinese Academy of Sciences, Beijing, China. Right: SpiralTape on Tabletop exhibited in the seventh-floor exhibition room in STM.

TABLE I
3 × 3 BETWEEN SUBJECTS FACTORIAL DESIGN.

		Video summarization patterns		
		SpiralTape (ST)	Tapestries (TS)	Rectangular tiles (RT)
Training time	No training (T0)	Using ST with T0	Using TS with T0	Using RT with T0
	15-min training (T15)	Using ST with T15	Using TS with T15	Using RT with T15
	30-min training (T30)	Using ST with T30	Using TS with T30	Using RT with T30

touch-capable devices in the Software Technology Museum located at the Institute of Software, at the Chinese Academy of Sciences, Beijing, China (Figure 8). During a one-week exhibition over the summer holidays in 2015, the majority of young visitors showed great interest in SpiralTape and interacted with it in high spirits. More details can be found in supplemental demo video.

A. Experimental Design

This study was conducted using a 3 × 3 between-subjects factorial design (Table I). 81 participants (45 females and 36 males) took part in this experiment and were divided into nine groups of equal size. Their ages ranged from 19 to 29 and all of them were at the college level. There was no significant difference in skills of computer operation and video editing among all participants, according to their self-reports and our observations.

Video summarization patterns for baseline evaluation. We compared SpiralTape with two baseline summarization schemes in linear patterns (Figure 1). One is the video tapestries [2] which uses a line-plot layout to represent the video timeline at a linear scale. To better fill up the available screen area, we broke the line of video tapestries at the screen border and tiled them into multiple lines to generate a second pattern named *rectangular tiles*. We also incorporated the user interface presented in Section VI for SpiralTape into the tapestries and rectangular tiles.

Training schedule. We were interested in the learnability of each video summarization pattern and designed three levels of training schedules:

- Schedule 1 (no training time). There was no training time before starting the evaluation session.
- Schedule 2 (15-minute training time). Before starting the evaluation session, each group was trained to use one video summarization pattern for approximately fifteen minutes.

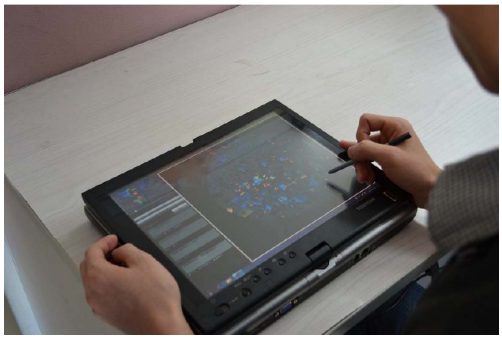


Fig. 9. Experimental environment using a Toshiba Tablet PC (Intel(R) Core(TM)2 CPU 2.00GHz) running Windows 7.

- Schedule 3 (30-minute training time). Before starting the evaluation session, each group was trained to use one video summarization pattern for approximately thirty minutes.

The detailed training procedure is presented in the Supplementary Material.

B. Tasks and Videos

Experimental tasks. Video summarization aims to provide a concise form so that a user can better grasp the content of a long video clip in a short time, and enable more efficient content indexing and access. In this study, we designed two tasks to evaluate and compare the effectiveness and efficiency of each video summarization pattern from these two aspects.

The first task was to understand video contents using one of the three video summarization patterns (Figure 1). For each video clip, there were four questions to answer in the first task. Participants were asked to select one correct answer from three candidates for each question. Two examples of these questions were "*Who brought a gift for Mike after the Mary Alice's funeral?*" and "*What did Bunny shoot at?*"

The second task was to locate some specific scenes that contain specified visual information in the video summarization. Similarly, participants were required to answer four questions and only one answer was correct for each question. Two examples were "*Find the scene where Lynette's three sons are playing in the swimming-pool*" and "*Find the scene in which Bunny was skipping rope*".

Video clips. Three video clips were extracted from the movies *The King of Milu* (2009), *Big Buck Bunny* (2008) and *Desperate Housewives* (2004). The clip lengths ranged from 15 to 20 minutes. The summarization of *The King of Milu* was used during the training process, where the other two summarizations were used in the formal test. All tasks were performed on a Toshiba Tablet PC (Intel(R) Core(TM)2 CPU 2.00GHz) running Windows 7 (Figure 9). All participants confirmed that they had not watched these selected videos previously.

C. Experimental Procedure

Upon arrival, participants were asked to sign a consent document. Several questions were designed to capture the

TABLE II
MEANS AND STANDARD DEVIATIONS FOR THE TTCS (TIME TO COMPLETE) AND ACCS (ACCURACY) OF THE FIRST AND SECOND TASKS.

Pattern		SpiralTape			Rectangular tiles			Tapestries		
Training time		0 mins	15 mins	30 mins	0 mins	15 mins	30 mins	0 mins	15 mins	30 mins
First Task	TTC (mins)	34.67 (2.45)	22.44 (1.67)	14.44 (1.74)	32.44 (2.79)	24.89 (2.67)	20.89 (3.79)	34.11 (2.47)	28.44 (2.19)	25.67 (3.5)
	ACC (%)	80.56 (11.2)	86.11 (11.6)	93.06 (6.59)	77.78 (10.4)	83.33 (10.8)	87.5 (12.5)	79.17 (12.5)	80.56 (12.7)	88.89 (7.5)
Second Task	TTC (mins)	14.11 (0.93)	9.11 (1.45)	5.11 (1.45)	15.11 (1.05)	10.67 (1)	8.67 (1.41)	15.56 (1.33)	12.22 (1.56)	10.44 (1.67)
	ACC (%)	76.39 (11.6)	84.72 (8.33)	91.67 (6.25)	79.17 (6.25)	86.11 (7.51)	88.89 (11.6)	77.78 (10.42)	83.33 (8.84)	87.5 (6.25)

participants' demographic information (e.g., age, gender, educational level), computer skills, and video editing experience. Before the training session, all participants watched a demo video that described how to use the operations provided in the user interfaces. During later training (if any), participants learned to use these interaction operations in a step-by-step manner.

- Three groups of users performed the first and second tasks with SpiralTape. Each group received one of the three training schedules (no training, 15-min and 30-min training time).
- Three groups of users performed both tasks with the rectangular tiles. Each group received one of the three training schedules.
- Three groups of users performed both tasks with the tapestries. Each group received one of the three training schedules.

For each video clip, participants answered the four questions for the first task and the four questions for the second task, followed by a short break. Participants were asked to complete both tasks as fast and accurately as possible. The time to complete (TTC) the first and second tasks (in minutes) with three summarization patterns, as well as the accuracy (ACC) of answers for both tasks, were recorded for each participant. During the process of both tasks, the interaction behaviors of each participant were captured. The number of occurrences in behavior is a basic behavioral recording type in applied behavior analysis [32]. We calculated the frequency of operations (in percentages) for each type of interaction behavior.

D. Experimental Results

A multivariate analysis of variance was performed on the collected data. Significant findings were followed up to assess the magnitude of the differences in performance among three video summarization patterns at each level of training schedule. The means and standard deviations for the TTC and ACC of both tasks are provided in Table II for each summarization pattern at each level of the training schedule. The average frequency of each interaction operation during the completion of both tasks is reported in Table III.

1) Performance in the first task: A significant interaction between the video summarization pattern and training schedule was revealed for the TTC of the first task ($F(4, 72) = 11.8$, $p < 0.001$) (see Figure 10). The main effects of the summarization pattern ($F(2, 72) = 29.6$, $p < 0.001$) and the training schedule ($F(2, 72) = 174.1$, $p < 0.001$) were significant for this measure.

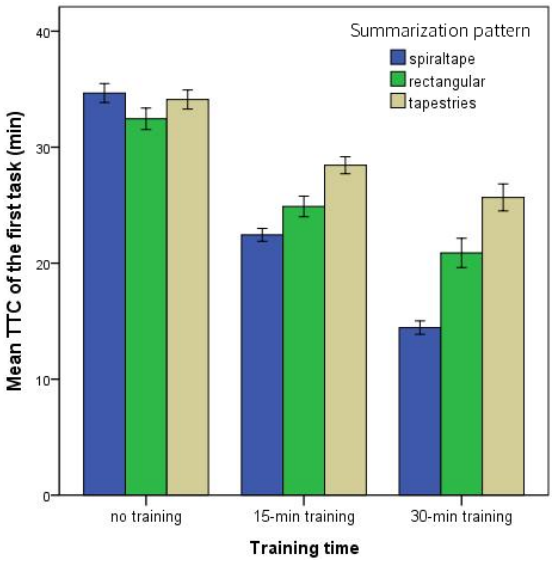


Fig. 10. A significant interaction between video summarization patterns and training schedules for the TTC of the fist task. The bars indicate ± 1 standard error.

Pair-wise comparisons showed no significant differences in TTCs among the three summarization patterns for participants who received no training before the evaluation session. In contrast, after a 15-minute training session, participants using SpiralTape spent the least time understanding the video content and answer the questions (see Table IV):

- There was a significant difference in the mean TTC of the first task between SpiralTape ($M = 22.44$ min) and rectangular tiles ($M = 24.89$ min) ($p < 0.05$).
- There was a significant difference in mean TTC of the first task between rectangular tiles ($M = 24.89$ min) and tapestries ($M = 28.44$ min) ($p < 0.01$).

The differences in TTCs among three summarization patterns became more evident when participants received 30 minutes of training:

- There was a significant difference in mean TTC of the first task between SpiralTape ($M = 14.44$ min) and rectangular tiles ($M = 20.89$ min) ($p < 0.001$).
- There was a significant difference in mean TTC of the first task between rectangular tiles ($M = 20.89$ min) and tapestries ($M = 25.67$ min) ($p < 0.01$).

TABLE III
THE AVERAGE FREQUENCY OF EACH INTERACTION OPERATION IN BOTH TASKS.

Interaction operation	Pattern / Training time	SpiralTape 0 mins	SpiralTape 15 mins	SpiralTape 30 mins	Rectangular tiles 0 mins	Rectangular tiles 15 mins	Rectangular tiles 30 mins	Tapestries 0 mins	Tapestries 15 mins	Tapestries 30 mins
Common type	Zoom in	32.65%	21.73%	15.29%	32.76%	30.41%	33.24%	36.10%	32.91%	33.08%
	Zoom out	27.15%	15.89%	5.20%	31.07%	30.79%	8.54%	27.58%	29.01%	7.46%
	Clicking ROI	30.17%	31.64%	37.61%	32.13%	30.92%	37.51%	30.93%	29.63%	37.11%
	Total	**89.97%**	**69.26%**	**58.10%**	**95.97%**	**92.12%**	**79.28%**	**94.61%**	**91.56%**	**77.65%**
Advanced type	Annotation	2.70%	10.42%	13.16%	1.80%	4.22%	11.43%	1.73%	4.54%	10.14%
	Hyperlink	1.51%	10.67%	8.41%	1.27%	2.86%	8.38%	2.05%	3.03%	10.88%
	Splitter and joiner	3.99%	6.10%	14.06%	0.64%	0.37%	0.60%	0.76%	0.50%	0.75%
	Visual percent	1.83%	3.56%	6.27%	0.31%	0.42%	0.30%	0.86%	0.37%	0.59%
	Total	**10.03%**	**30.74%**	**41.90%**	**4.03%**	**7.88%**	**20.72%**	**5.39%**	**8.44%**	**22.35%**

TABLE IV
PAIR-WISE COMPARISONS FOR THE TTCS OF THE FIRST AND SECOND TASKS. CI FOR CONFIDENCE INTERVAL.

Pair-wise comparisons		No training		15-min training		30-min training	
		p-value	i-j (95% CI)	p-value	i-j (95% CI)	p-value	i-j (95% CI)
TTC of the first task (min)	SpiralTape (i) vs. Rectangular tiles (j)	Not significant		$< .05$	-2.44 $(-4.6, -.29)$	$< .001$	-6.44 $(-9.5, -3.39)$
	SpiralTape (i) vs. Tapestries (j)	Not significant		$< .001$	-6.0 $(-8.15, -3.85)$	$< .001$	-11.22 $(-14.3, -8.16)$
	Rectangular tiles (i) vs. Tapestries (j)	Not significant		$< .01$	-3.56 $(-5.71, -1.4)$	$< .01$	-4.78 $(-7.84, -1.72)$
TTC of the second task (min)	SpiralTape (i) vs. Rectangular tiles (j)	Not significant		$< .05$	-1.56 $(-2.88, -.23)$	$< .001$	-3.56 $(-5.03, -2.08)$
	SpiralTape (i) vs. Tapestries (j)	Not significant		$< .001$	-3.11 $(-4.44, -1.79)$	$< .001$	-5.33 $(-6.81, -3.86)$
	Rectangular tiles (i) vs. Tapestries (j)	Not significant		$< .01$	-1.56 $(-2.88, -.23)$	$< .01$	-1.78 $(-3.25, -.3)$

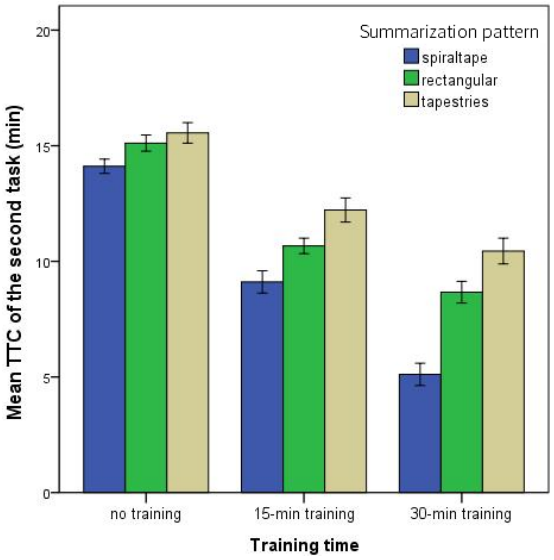


Fig. 11. A significant interaction between video summarization pattern and training schedule for the TTC of the second task. The bars indicate ± 1 standard error.

In terms of the mean accuracy of answers, the main effect of the training schedule was significant ($F(2, 72) = 6.63$, $p < 0.01$). As users received more training experience, they achieved a higher accuracy in understanding the video content. In contrast, there was no significant difference in accuracy among the three summarization patterns, regardless of the level of training users received. The results were as follows:

- No training: 80.56% in SpiralTape, 77.78% in rectangular tiles and 79.17% in tapestries.
- Fifteen minutes of training: 86.11% in SpiralTape, 83.33% in rectangular tiles and 80.56% in tapestries.
- Thirty minutes of training: 93.06% in SpiralTape, 87.5% in rectangular tiles and 88.89% in tapestries.

2) Performance in the second task: We observed a significant interaction between video summarization pattern and training schedule for the TTC of the second task ($F(4, 72) = 5.11$, $p = 0.001$) (see Figure 11). Both main effects of the summarization pattern ($F(2, 72) = 41.52$, $p < 0.001$) and training schedule ($F(2, 72) = 179.6$, $p < 0.001$) were significant for this measure.

Pair-wise comparisons showed no significant differences in TTCs among the three summarization patterns when participants received no training before the evaluation session. However, after a fifteen-minute training session, participants using SpiralTape spent the least time locating the specified scenes (see Table IV):

- There was a significant difference in mean TTC of the second task between SpiralTape ($M = 9.11$ min) and rectangular tiles ($M = 10.67$ min) ($p < 0.05$).
- There was a significant difference in mean TTC of the second task between rectangular tiles ($M = 10.67$ min) and tapestries ($M = 12.22$ min) ($p < 0.05$).

The differences in TTCs among three summarization patterns became more evident when participants received 30 minutes of training:

- There was a significant difference in mean TTC of the second task between SpiralTape ($M = 5.11$ min) and rectangular tiles ($M = 8.67$ min) ($p < 0.001$).

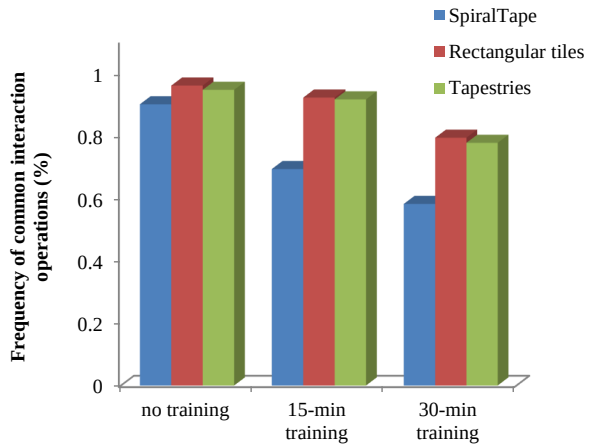


Fig. 12. The average frequency of common interaction operations at each level of training schedule. The frequency of interaction operation in common type and advanced type was equal to 1.

- There was a significant difference in mean TTC of the second task between rectangular tiles ($M = 8.67$ min) and tapestries ($M = 10.44$ min) ($p < 0.05$).

In terms of the mean accuracy of answers, the main effect of the training schedule was significant ($F(2, 72) = 11.8$, $p < 0.001$). As users received more training experience, they achieved a higher accuracy when locating the specified scenes. However, there was no significant difference among the three summarization patterns, regardless of the level of training users received. The results were as follows:

- No training: 76.39% in SpiralTape, 79.17% in rectangular tiles and 77.78% in tapestries.
- Fifteen minutes of training: 84.72% in SpiralTape, 86.11% in rectangular tiles and 83.33% in tapestries.
- Thirty minutes of training: 91.67% in SpiralTape, 88.89% in rectangular tiles and 87.5% in tapestries.

3) Interaction behavior: There are seven interaction operations provided by this user interface that can be classified into two types:

- Common types include the three basic video viewing and browsing operations: zoom in, zoom out, view video by clicking an ROI.
- Advanced types include the four new operations proposed in Section VI: annotations, hyperlinks, splitter and joiner, and visual percent of scenes.

The results showed that as participants received more training time, they used a greater number of advanced interaction operations and had a higher interaction efficiency (see Table III and Figure 12). The results were as follows:

- No training: the total number of interaction operations was 103.12 (89.97% common types) in SpiralTape, 104.78 (95.97% common types) in rectangular tiles and 103.11 (94.61% common types) in tapestries.
- Fifteen minutes of training: the total number of interaction operations was 87.44 (69.26% common types) in SpiralTape, 89.5 (92.12% common types) in rectangular tiles and 88.11 (91.56% common types) in tapestries.
- Thirty minutes of training: the total number of interaction operations was 72.67 (58.1% common types) in SpiralTape, 72.87 (79.28% common types) in rectangular tiles and 74.57 (77.65% common types) in tapestries.

Without any training experience, the frequency of common interaction operations was similar for the three summarization patterns. Notable differences between SpiralTape and the other two summarization patterns ($> 20\%$) were observed with increasing training time. Participants with SpiralTape tended to use more advanced interaction operations to enhance content indexing and access.

E. More Experiments

1) Training efficiency: We further examine the effect of training experience on the performance. As shown in Figures 10 and 11, the differences between two successive training schedules decreased and learning curves (i.e., a body of knowledge was learned over time) became closer to flat for the rectangular tiles and tapestries. In contrast, the learning curve of SpiralTape showed a rapid increase in learning with additional training time. As a result, this section describes an extended version of the previous user study that investigated the efficiency of further training on interaction performance with SpiralTape.

This extended session involved 18 participants (8 females and 10 males). Their ages ranged from 24 to 40, and all participants were currently in college or had graduated from college. Confounding effects of prior computer skills and video editing experience were controlled. Participants were divided into two groups of equal size. One group was trained to use SpiralTape for approximately forty-five minutes, whereas the other group received training instructions for one hour before starting the evaluation session. The experimental settings and procedure were identical to those in the previous study, and participants' TTC in the first and second tasks were recorded.

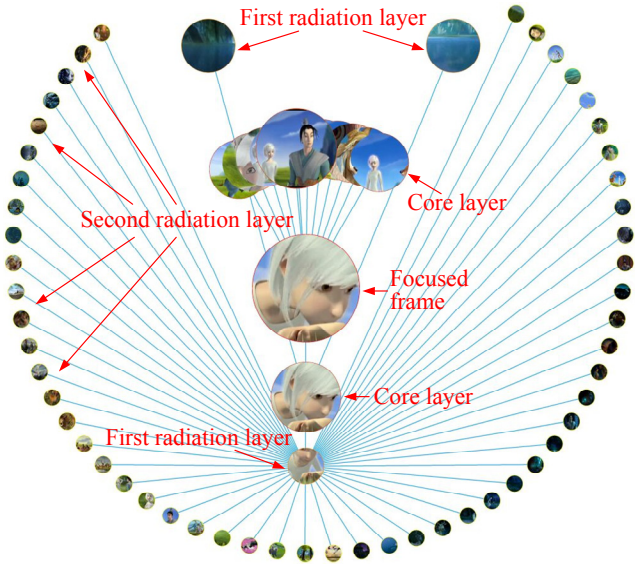


Fig. 13. A radial method [24] uses three layers (one core layer and two radiation layers) to visualize video structures.

A one-way analysis of variation (ANOVA) was performed, and the training schedule was treated as an independent variable with three levels: 30 minutes of training, 45 minutes of training, and 60 minutes of training.

- There was no significant difference in mean TTC of the first task between the 30-minute ($M = 14.44$ min) and 45-minute training schedules ($M = 13.78$ min).
- There was no significant difference in mean TTC of the first task between the 45-minute ($M = 13.78$ min) and 60-minute training schedule ($M = 13.44$ min).
- There was no significant difference in mean TTC of the second task between the 30-minute ($M = 5.11$ min) and 45-minute training schedule ($M = 5.33$ min).
- There was no significant difference in mean TTC of the second task between the 45-minute ($M = 5.33$ min) and 60-minute training schedule ($M = 4.78$ min).

2) Comparison with a radial video visualization: We also compared SpiralTape with a state-of-the-art radial video visualization method [24]. This radial method visualizes video structures by arranging frames in three layers (Figure 13):

- Core layer. Keyframes are shown in this layer and equally distributed around the circle center.
- First radiation layer. Representative frames that have content similar to the current focused key frame are shown in this layer.
- Second radiation layer. Corresponding details about each video shot are shown in this layer.

In both radiation layers, frames are rendered in a uniform radiation centered at the focused frame. To reflect the temporal structure, frames are visualized in the chronological order from inside to outside the circle in a clockwise direction. The interaction operations designed in [24] included retrieval of any interesting frame on demand, manually adjusting frame scaling and interactively controlling the frame drawing process.

The 18 participants who attended the extended session were also asked to watch a demo video illustrating the radial method [24]. Half of the participants watched the demo video before the the extended session and the other half watched the demo video after the extended session. Finally, they completed the questionnaire shown in Figure 14.

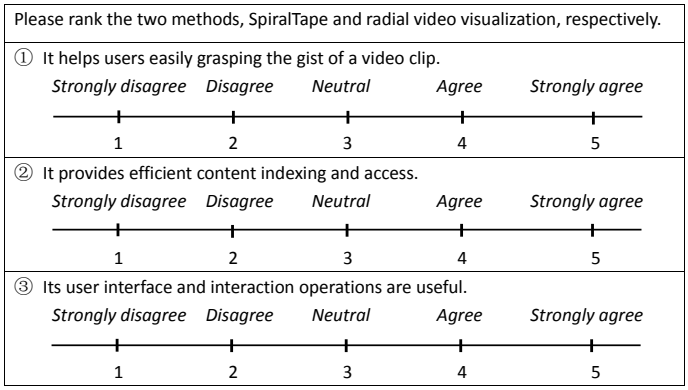


Fig. 14. A questionnaire presented to participants after they completed the extended session and watched a demo video illustrating the radial method [24].

A paired t-test was performed and the results showed that:

- There was a significant difference in the mean evaluation scores of Question 1 (it helps users easily grasping the gist of a video clip) between SpiralTape ($M = 4.33$) and the radial method ($M = 3.61$), $p < 0.05$.
- There was no significant difference in the mean evaluation scores of Question 2 (it provides efficient content indexing and access) between SpiralTape ($M = 4.11$) and the radial method ($M = 3.78$).
- There was a significant difference in the mean evaluation scores of Question 3 (its user interface and interaction operations are useful) between SpiralTape ($M = 4.28$) and the radial method ($M = 3.50$), $p < 0.01$.

VIII. Conclusion

In this paper, a new video summarization *SpiralTape* is proposed for facilitating users in understanding and indexing video content quickly and efficiently. Several design principles, including compactness, continuity, efficient overview and interactivity, are considered when designing SpiralTape. To implement these design principles, we preprocess video clips into hierarchical structures {events, scenes, shots, keyframes} and extract ROIs from the keyframes. When a user interactively browses SpiralTape in a personalized manner, these ROIs are arranged in a compact spiral layout that supports smooth transition and hierarchical emergence. A user interface with gesture operations is provided in SpiralTape to help users browse video content naturally and fluidly. An elaborate user study demonstrated that after training for some time, SpiralTape outperformed two representative linear video summarization methods and a state-of-the-art radial video visualization.

Acknowledgment

The authors thank the editor and reviewers for their valuable comments, which help improve this paper. This work was done when the first author visited Intelligence Engineering Lab, Institute of Software. Y-J Liu and C-X Xia contributed equally to this paper and are co-corresponding authors.

References

[1] G. P. Nguyen and M. Worring, "Optimization of interactive visual-similarity-based search," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 1, pp. 7:1–7:23, 2008.

[2] C. Barnes, D. B. Goldman, E. Shechtman, and A. Finkelstein, "Video tapestries with continuous temporal zoom," in *ACM SIGGRAPH '10*, 2010, pp. 89:1–89:9.

[3] C. D. Correa and K.-L. Ma, "Dynamic video narratives," in *ACM SIGGRAPH '10*, 2010, pp. 88:1–88:9.

[4] C. Ma, Y.-J. Liu, H. Yang, D. Teng, and G. Dai, "Sketch-based annotation and visualization in video authoring," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1153–1165, 2012.

[5] Y.-J. Liu, C.-X. Ma, Q. Fu, X. Fu, S.-F. Qin, and L. Xie, "A sketch-based approach for interactive organization of video clips," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 1, pp. 2:1–2:21, 2014.

[6] C. Tominski and H. Schumann, "Enhanced interactive spiral display," in *The Annual SIGRAD Conference, Special Theme: Interaction*. Linköping University Electronic Press, 2008, pp. 53–56.

[7] M. Weber, M. Alexa, and W. Müller, "Visualizing time-series on spirals," in *IEEE Symposium on Information Visualization (INFOVIS '01)*, 2001, pp. 7–14.

[8] E. Graells and A. Jaimes, "Lin-spiration: Using a mixture of spiral and linear visualization layouts to explore time series," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI '12)*, 2012, pp. 237–240.

[9] J. Lin, E. Keogh, and S. Lonardi, "Visualizing and discovering non-trivial patterns in large time series databases," in *IEEE Symposium on Information Visualization (INFOVIS '05)*, 2005, pp. 61–82.

[10] S. Lu, Z. Wang, T. Mei, G. Guan, and D. D. Feng, "A bag-of-importance model with locality-constrained coding based feature learning for video summarization," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1497–1509, 2014.

[11] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake, "Digital tapestry," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '05)*, 2005, pp. 589–596.

[12] T. Mei, B. Yang, S.-Q. Yang, and X.-S. Hua, "Video collage: presenting a video sequence using a single image," *The Visual Computer*, vol. 25, no. 1, pp. 39–51, 2009.

[13] M. Yu, Y.-J. Liu, S.-J. Wang, Q. Fu, and X. Fu, "A pmj-inspired cognitive framework for natural scene categorization in line drawings," *Neurocomputing*, vol. 173, p. 2041C2048, 2016.

[14] Y.-J. Liu, M. Yu, Q. Fu, W. Chen, Y. Liu, and L. Xie, "Cognitive mechanism related to line drawings and its applications in intelligent processing of visual media: a survey," *Frontiers of Computer Science*, vol. 10, no. 2, pp. 216–232, 2016.

[15] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, "Video manga: Generating semantically meaningful video summaries," in *Proceedings of the 7th ACM International Conference on Multimedia (MM '99)*. ACM, 1999, pp. 383–392.

[16] M. A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," in *IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998, pp. 61–70.

[17] A. Shamir and O. Sorkine, "Visual media retargeting," in *ACM SIGGRAPH ASIA 2009 Courses*, 2009, pp. 11:1–11:13.

[18] Y.-J. Liu, X. Luo, Y.-M. Xuan, W.-F. Chen, and X.-L. Fu, "Image retargeting quality assessment," *Computer Graphics Forum (Eurographics 2011)*, vol. 30, no. 2, pp. 583–592, 2011.

[19] Y. Liang, Y.-J. Liu, and D. Gutierrez, "Objective quality prediction of image retargeting algorithms," *IEEE Transactions on Visualization and Computer Graphics*, DOI: 10.1109/TVCG.2016.2517641, 2016.

[20] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.

[21] J. Yang, M. O. Ward, and E. A. Rundensteiner, "Interring: An interactive tool for visually navigating and manipulating hierarchical structures," in *IEEE Symposium on Information Visualization (INFOVIS '02)*, 2002, pp. 77–84.

[22] G. M. Draper, Y. Livnat, and R. F. Riesenfeld, "A survey of radial methods for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 5, pp. 759–776, 2009.

[23] J. V. Carlis and J. A. Konstan, "Interactive visualization of serial periodic data," in *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology (UIST '98)*, 1998, pp. 29–38.

[24] Y. Wu, S. Wang, H. Wang, Q. Li, H. Jiang, and Y. Zou, "A total variation-based hierarchical radial video visualization method," *Journal of Visualization*, vol. 18, no. 2, pp. 255–267, 2015.

[25] L. Herranz, J. Ćalić, J. M. Martínez, and M. Mrak, "Scalable comic-like video summaries and layout disturbance," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1290–1297, 2012.

[26] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 1–13, 2000.

[27] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, 2010.

[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[29] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972–976, 2007.

[30] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *In Proc. IEEE Inter. Joint Conf. Artificial intelligence (IJCAI81)*, 1981, pp. 674–679.

[31] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems (NIPS '07)*, 2007, pp. 545–552.

[32] J. L. Michael, *Concepts and Principles of Behavior Analysis*. Association of Behavior Analysis; Revised Edition, 2004.