A Sketch-Based Approach for Interactive Organization of Video Clips

YONG-JIN LIU, Tsinghua University
CUI-XIA MA, QIUFANG FU, and XIAOLAN FU, Chinese Academy of Sciences
SHENG-FENG QIN, Northumbria University
LEXING XIE, The Australian National University

With the rapid growth of video resources, techniques for efficient organization of video clips are becoming appealing in the multimedia domain. In this article, a sketch-based approach is proposed to intuitively organize video clips by: (1) enhancing their narrations using sketch annotations and (2) structurizing the organization process by gesture-based free-form sketching on touch devices. There are two main contributions of this work. The first is a sketch graph, a novel representation for the narrative structure of video clips to facilitate content organization. The second is a method to perform context-aware sketch recommendation scalable to large video collections, enabling common users to easily organize sketch annotations. A prototype system integrating the proposed approach was evaluated on the basis of five different aspects concerning its performance and usability. Two sketch searching experiments showed that the proposed context-aware sketch recommendation outperforms, in terms of accuracy and scalability, two state-of-the-art sketch searching methods. Moreover, a user study showed that the sketch graph is consistently preferred over traditional representations such as keywords and keyframes. The second user study showed that the proposed approach is applicable in those scenarios where the video annotator and organizer were the same person. The third user study showed that, for video content organization, using sketch graph users took on average 1/3 less time than using a mass-market tool Movie Maker and took on average 1/4 less time than using a state-of-the-art sketch alternative. These results demonstrated that the proposed sketch graph approach is a promising video organization tool.

Categories and Subject Descriptors: D.2.10 [**Software Engineering**]: Design—*Representation*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Sketching interface, video organization, sketch annotation, context-aware recommendation

ACM Reference Format:
Yong-Jin Liu, Cui-Xia Ma, Qiufang Fu, Xiaolan Fu, Sheng-Feng Qin, and Lexing Xie. 2014. A sketch-based approach for interactive organization of video clips. ACM Trans. Multimedia Comput. Commun. Appl. 11, 1, Article 2 (August 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/2645643

This work is supported by The Natural Science Foundation of China under grants 61322206, 61173058, the 973 program of China under grant 2011CB302202, the 863 Program of China under grant 2012AA011801, and the Tsinghua University Initiative Scientific Research Program, under grant 20131089252.
Authors' addresses: Y.-J. Liu (corresponding author), TNList, Department of Computer Science and Technology, Tsinghua University, China; email: liuyongjin@tsinghua.edu.cn; C.-X. Ma, State Key Lab of Computer Science, Institute of Software, and Q. Fu and X. Fu, State Key Lab of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, China; S.-F. Qin, Department of Design, Northumbria University, UK; L. Xie, Research School of Computer Science, The Australian National University, Australia.

DOI: http://dx.doi.org/10.1145/2645643

1. INTRODUCTION

State-of-the-art video capturing techniques make it easy to acquire a large set of video clips in daily life. For companies or home users, a frequently encountered problem is how to efficiently organize these video clips and reuse them later for some particular tasks. Accordingly, there is an increasing demand for a new organization method for video clips that is easy to use and has a natural interaction with users. In order to design and provide a nice organization tool for the user, the following two questions have to be answered: (1) How do we represent the organization information in a structural form? (2) How do we make sure that the structural form has a good capacity for summarization and visualization, not only for a single video clip but also for interrelations among different video clips? To address these questions, two closely related works are video summarization and multimedia authoring, which are summarized in the next section.

In state-of-the-art video processing software such as Adobe Premiere Pro and Microsoft Movie Maker, the user can browse a video clip along its timeline represented by a sequence of keyframes. For organizing multiple video clips, the user needs to frequently scroll the timelines right-and-left, as well as locate and operate on some frames from different video clips, resulting in excessive operations on pressing buttons and selecting in menus with mouse and keyboard input. As a comparison, sketching is more natural and intuitive for interaction with touch devices. In this article, we present a sketch-based approach for interactive organization of video clips. The novelty in the presented approach includes two aspects.

—*Intuitive video content organization utilizing efficient personalized video annotation by sketches.* Instead of traditional keyframes and keywords, we use line drawings made by users' sketching as concise personalized annotations. We show that, either users' simple sketches or simple modification on these sketches automatically converted from keyframes is intuitive and fast, and can efficiently reduce the gap between high-level semantics and low-level image features. To structure the organization process, we propose a sketch graph wherein gesture-based free-form sketching is used to efficiently represent the user's organization intent. The gesture-based sketching organization is also consistent with the sketch annotation to make a coherent sketching system.

—*Context-aware sketch recommendation.* It is desired that, during video organization, a common user only needs to draw some simple shapes to represent her/his intent and then the related vido clips with nice sketch annotations can be automatically recommended. To meet this requirement, we propose a context-aware sketch recommendation method that can efficiently capture the user's intent and recommend related sketch annotations in the database based on the user's simple drawings. Also studied is the scalability of the proposed sketch recommendation that makes the proposed sketching system suitable for a large-scale video database.

We evaluated the sketch representation (both generated sketch annotations and sketch graphs) for video organization in three user studies in Sections 4.3, 4.4, and 4.5, and found that: (1) it had a better performance in visualization of organization intent than keywords and keyframes, (2) it was applicable in those scenarios where the video annotator and organizer were the same person, and (3) it was more efficient and took less time in video content organization compared to a mass-market tool Movie Maker and a state-of-the-art sketch alternative [Ma et al. 2012]. We also evaluated the proposed context-aware sketch recommendation in Sections 4.1 and 4.2, and showed that it is more efficient in terms of accuracy and scalability than the state-of-the-art edgel index method [Cao et al. 2011] and the visual word method [Ma et al. 2012], and naturally scales to a large database with more than 1 million sketches.

In the remainder of the article, Section 2 summarizes the related work. The proposed sketch-based approach with detailed implementation is presented in Section 3. Section 4 presents the experiments and user studies. Finally Section 5 concludes.

2. RELATED WORK

A few works have addressed the video organization problem from different perspectives. In Das and Liou [1998], each video clip is automatically processed by: (1) detecting the shot boundary, (2) grouping similar shots into shot groups, and (3) generating a video table of contents. The purpose of video organization in Das and Liou [1998] is to achieve a content-based indexing of video archives. In Zhang and Lu [2002], a video clip is organized into four layers {frame, shot, episode, video program} to facilitate indexing, browsing, and querying. Similarly, in Zhu et al. [2005] a video clip is organized into a hierarchy of video contents {keyframe, shot, group, scene, video} with which hierarchical video browsing and retrieval are integrated for efficient video access. These works focused on low-level image and video feature processing such as shot boundary detection and grouping video clips. As a comparison, the presented work in this article focuses on representing and visualizing the organization structure among a set of video clips. In order to enable a user to efficiently build such an organization structure, different from previous works that mainly use the WIMP interaction style, the interaction style of sketching is considered in this article. Next, we summarize the related work on video summarization, video annotation, sketching interfaces, sketch retrieval/recognition, and multimedia authoring.

Video summarization. Video summarization is one of the efficient methods that allow users to quickly capture video contents. Most video summarization methods extract a collection of keyframes and display them in some elegant forms. Broadly speaking, the presentation of extracted keyframes can be classified into two forms [Truong and Venkatesh 2007]: storyboard or dynamic slideshow. A state-of-the-art dynamic slidshow work [Correa and Ma 2010; Zhang et al. 2013] emphasized the character's motion in video data and composed a matted motion foreground on a panoramic background. However, the drawback of dynamic slideshow forms is that the user can only view them passively and may lose the context information between keyframes. For the storyboard form, both photo-realistic collage [Mei et al. 2008] and a sketch-like schemas [Ma et al. 2012] can be used. For the purpose of video clip organization, sketch-like schemas using line drawings, flowchart symbols, and texts are better for visualizing and editing the schematic storyboard layout, without the need to compute the fine details in a realistic composed image, such as consistent texture and lighting.

Video annotation. Borgo et al. [2011] argue that a good video summarization should not intend to provide a fully automatic solution, but to provide an efficient tool for communicating messages among people. Compared to the automatic video summarization by shot detection and grouping based on low-level image features, semantic annotation in video can provide valuable information for video content understanding and help the user to generate desired video summarization in an interactive way. Video annotation can be either automatic [Wang and Hua 2011b] or interactive [Ma et al. 2012]. In our study, due to the possible large variance and diversity of contents in video clips that need to be organized in an interactive process, we use an interactive video annotation as a preprocess that allows the user to input annotations for facilitating content understanding and drawing attention from others.

Sketches and sketching interface. When a user inputs semantic annotations in video, sketches such as hand-drawn figures, symbols (e.g., arrows and links), or handwritten texts are desired when using modern touch devices such as iPad, smartphones, and tablet PCs [Ma et al. 2012]. It is natural to provide a sketching interface for the generation of both sketch annotations and sketch-like schematic storyboards. Sketching has

been shown a useful tool to naturally express the design/organization intent of users [Rodgersa et al. 2000]. In the presented work on sketch-based organization of video clips, our study shows that, during the organization process, the sketching interface efficiently helps prevent users' thoughts from being interrupted by excessively switching among menus, buttons, and keyboard as with most commercial software like Adobe Premiere Pro and Microsoft Movie Maker.

Sketch retrieval and recognition. To reuse the knowledge in sketch annotations that users input in video clips, sketch retrieval and recognition are useful for organization. Since sketches only capture rough information of video contents, they are usually incomplete and undergo elastic deformation while remaining perceptually similar. Accordingly, three classes of techniques have been proposed. The first class segments sketches into meaningful pieces and then supports a partial matching for sketch recognition (e.g., [Ma et al. 2011; Sun et al. 2012]). The second class builds some intermediate shape signatures that can tolerate distortions between similar sketches, such as the edge index in Cao et al. [2011] and the circular histogram in Liu et al. [2013b]. The third class combines color information with sketches for supporting a large-scale image search [Sun et al. 2013; Wang and Hua 2011a]. All these previous works measure the similarity only between a single input sketch and a single sketch/image in the database. Note that, for sketch annotation, a video clip is usually represented by several sketches. In this article, different from these previous works, we propose a context-aware sketch recommendation in which, in addition to sketch similarity, the contexts of sketches in both video clips and sketched storyboards are also taken into consideration.

Multimedia authoring. Organization of video material is a key part in multimedia authoring [Bulterman and Hardman 2005]. Anecdote [Harada et al. 1996] is a classic multimedia authoring tool that provides various authoring styles to construct a scenario framework for video clips. The Anecdote tool supports an early design process by enabling users to collect a set of annotated storyboards, connect them with links, and arrange them in a time sequence. DEMAIS [Bailey et al. 2001] is another interactive multimedia storyboard that can help users design multimedia applications by using both ink strokes and textual annotations as an input design vocabulary. DEMAIS enables a user to quickly sketch behavioral design ideas and edit the behavior using an expressive visual language, similar to the sketch annotations and structure representation used in our approach. The key difference in our approach is that we use sketch annotation to capture the semantic meaning in video contents and that we propose a sketch graph to represent an organization structure which is intuitive to edit and modify during an interactive organization process.

3. SKETCH-BASED ORGANIZATION OF VIDEO CLIPS

The organization of video clips in the presented study for structuring and managing a set of video clips to meet a need or to pursue collective goals. In this section, we propose a sketch-based approach for video organization, with the following two characteristics.

—*Visualization of video contents by concise sketches.* Video is a typical form of visual information. At the fine granularity of pixel level, videos always exhibit superfluous information of shape, color, and texture. In this work we propose to use sketch annotations (Section 3.1) for a concise visualization of video shots and use a sketch graph (Section 3.2) for an efficient visualization of organization structures. Based on the sketch annotation and the sketch graph, a video organization is performed in a visualization-optimized manner that maps the sketch graph to the user's perception. Following the Gestalt law [Kanizsa 1979], here the user's perception means grouping

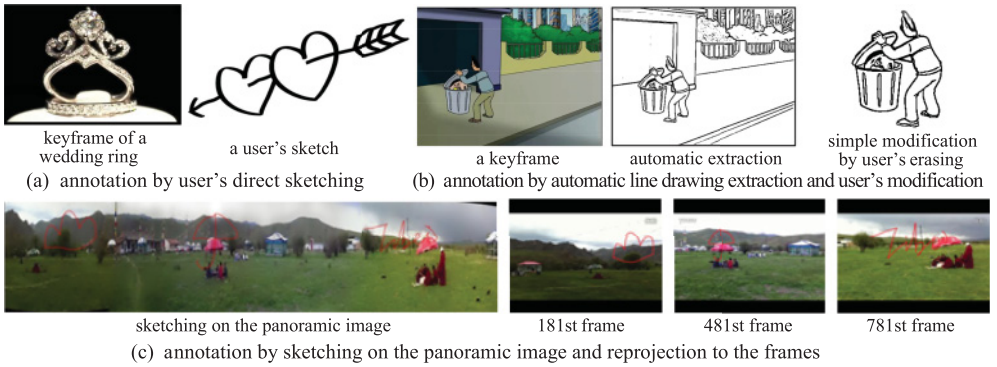


Fig. 1. Three annotation manners in the proposed method. All three examples in (a), (b), and (c) took only a few seconds in a modest interaction with a user and thus our personalized sketch annotations are efficient.

elementary perceptual elements (sketches) into larger structures (sketch graph) and understanding the relation among visual stimuli and their perceptions.

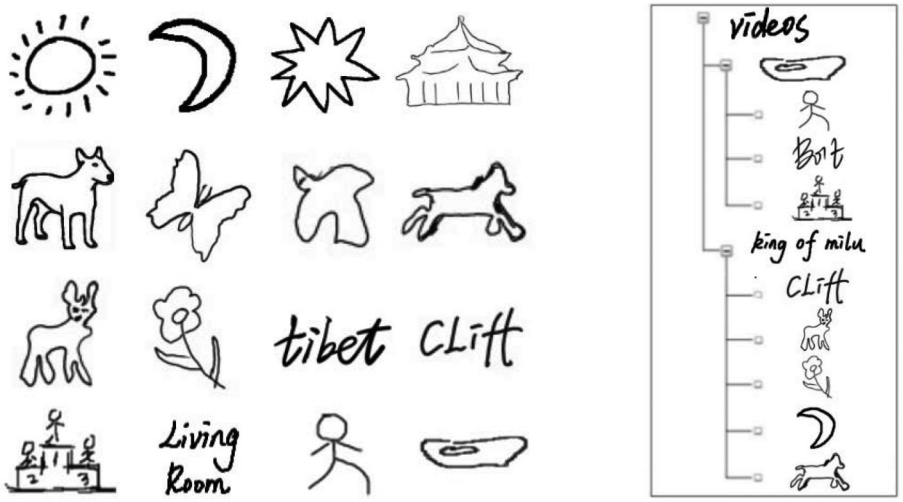
—*Knowledge organization using sketch graph and sketch recommendation.* To reflect users' organization intent, two kinds of knowledge are used in cognitive psychology [Best 1986]: declarative knowledge and procedural knowledge. Declarative knowledge refers to factual information that is static, which characterizes "knowing that". Procedural knowledge, on the other hand, refers to the knowledge of how to perform/operate a task, which characterizes "knowing how". In the proposed sketch graph model (Section 3.2), we use sketch nodes to represent declarative knowledge and sketching connections to represent procedural knowledge. Thus, the proposed sketch graph unifies two kinds of knowledge in a single and coherent model. To support an efficient knowledge reuse of sketch annotations in the video database, a context-aware sketch recommendation technique (Section 3.3) is proposed to facilitate users in efficiently retrieving sketch annotations by means of simple sketches.

3.1. Sketch Annotations

Different from previous works (e.g., [Moxley et al. 2010; Wang et al. 2009]) that use low-level image features to annotate video clips, our approach uses sketch annotations that can be efficiently obtained by either users' simple drawings (Figures 1(a), 1(c), and 2(a)) or a modest modification of nice sketches automatically extracted from keyframes (Figure 1(b)). The sketch annotations personalized by users' modest interactions can efficiently reduce the gaps between high-level semantics and low-level image features and better characterize the user's thought on the gist of the video content.

A video clip can be hierarchically decomposed into shots and frames. Accordingly, three sketching manners are provided in our system for sketch annotations.

(1) When a video clip is browsed frame by frame, the user can directly draw simple sketches on the frames. The user's personalized annotations can enrich and extend the video content from its low-level image features. For example, a user may draw two hearts pierced by an arrow on the frames of a diamond wedding ring to mean "love" (Figure 1(a)).
(2) Keyframes are extracted from a video clip and automatically converted into line drawings [Kang et al. 2007; Yu et al. 2014]. Then the user can modify these sketches by means of simple gesture operations. Modest modifications of existing line drawings can alleviate the users' burden of drawing complex sketches (Figure 1(b)).



(a) some simple sketches drawn by users in video annotation

(b) a hierarchy of sketches to represent video clips

Fig. 2. Some sketch annotations on video clips and video structure represented by sketch annotations, drawn by Yong-Jin Liu, Cui-Xia Ma © Yong-Jin Liu, Cui-Xia Ma.

(3) In a preprocessing stage, shots in the video clip are detected. If a shot taken is with a continuous camera motion, a panoramic image can be generated for this shot [Irani and Anandan 1998; Mei and Hua 2008]. Then the user can draw sketches on panoramic images (Figure 1(c)). The panoramic image can provide a large scale of static background that facilitates easy sketching of the user's personalized annotations.

After adding annotations, video clips are represented and retrieved based on the sketch annotations attached to them (Figure 2(b)). During the interactive organization process using a sketch graph (Section 3.2), these sketches can aid the indexing, management, and recommendation of related video clips, based on a context-aware sketch recommendation technique (Section 3.3).

3.2. Sketch Graph

Interactive organization of video clips is a design process. At an early design phase, a user may probably have only a rough idea of what she/he is looking for and what she/he wants. Free-hand sketching has been demonstrated as an efficient tool for communicating and recording rough ideas [Rodgersa et al. 2000]. On the other hand, an intuitive and concise visualization of organization information with video contents can help the user to efficiently edit and modify the organization. In our study, we propose a simple and concise form, called *sketch graph*, that uses a sketch-based storyboard as a visualization form and integrates connection lines to construct a narrative structure for organizing video clips.

A sketch graph consists of two parts: sketch nodes and sketching connections.

—*Sketch nodes of a sketch graph.* Each node is an annotated sketch representing a shot or a video clip (Figure 2(a)). The arrangement of nodes in an organization canvas gives an interactive storyboard.

—*Sketch connections among sketch nodes.* A connection between two nodes represents certain relations and a sketch connection can be added between any two nodes by the

(a) sequential and jump tracing (b) conditional (c) iterative

Fig. 3. Examples of different connection line types, drawings by Yong-Jin Liu © Yong-Jin Liu.

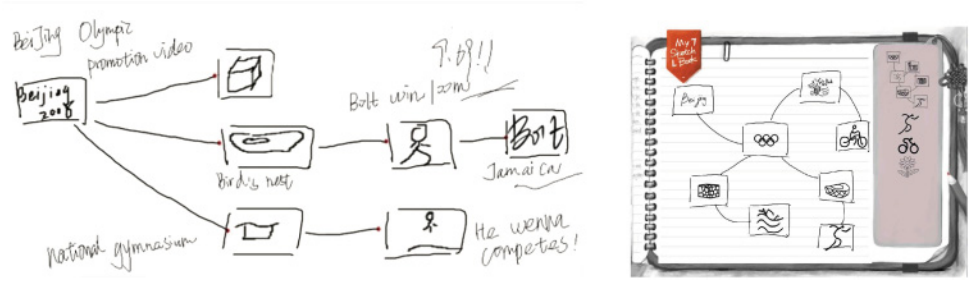


Fig. 4. Two examples of the sketch graph without (left) and with (right) the sketching interface. A snapshot of the sketching system is shown in Figure 7, drawings by Cui-Xia Ma © Cui-Xia Ma.

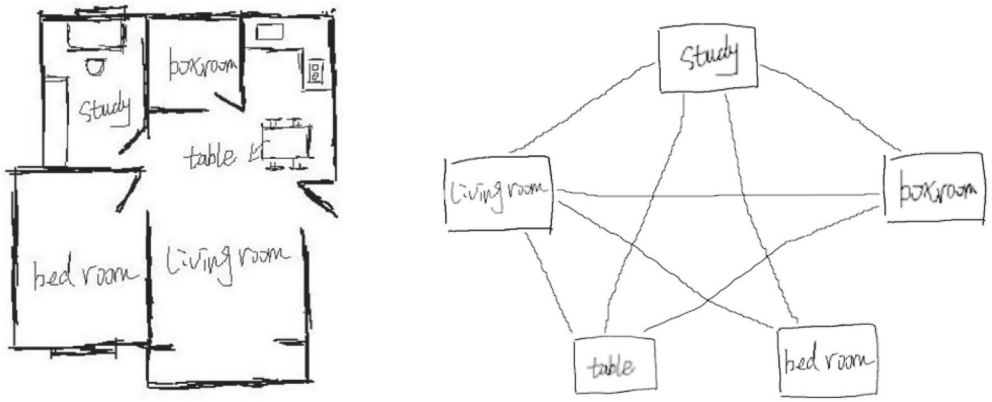
user (Figure 3). Different from most previous storyboards that usually use temporal relations, the sketch connections here can represent any relations, either physically (e.g., temporal or spatial orders) or logically (e.g., two video clips are both about sports).

Two examples of the sketch graph are shown in Figure 4. The sketch connections with sketching activity offer flexible control of the organization intent [Rodgersa et al. 2000] and can also easily depict nonlinear structures of the organization. Figure 3 shows several examples of different types of sketch connections mimicking the flowchart control, including sequential and jump tracing, as well as conditional and iterative connections. The sketching connection type can also be tailored to any particular application; for example, in a house exhibition, using sketch connections as a metaphor, a room layout can be converted into spatial adjacency relations (Figure 5).

Two kinds of knowledge, namely declarative and procedural, had been used for memory organization in cognitive psychology [Best 1986]. The sketch graph makes use of these two kinds of knowledge for organization as follows.

—*Declarative knowledge and sketch nodes.* Declarative knowledge is some factual information and often takes the form of a series of related facts. The sketch graph uses the sketch nodes to represent declarative knowledge. Rich semantic information can be conveyed with a single sketched drawing; as an old saying goes, "a good picture is worth a thousand words." Compared to the representation of keyframes and typed keywords, sketches behave more like a concept prototype in the human brain [Fu et al. 2013; Liu et al. 2013a].

—*Procedural knowledge and sketch connections.* Procedural knowledge consists of the skills in performing some tasks that is naturally dynamic. Procedural knowledge is often task dependent and thus is well known to be less general than declarative knowledge [Best 1986]. The sketch graph uses the sketch connections to implicitly represent some procedural knowledge. For example, sketch connections can



(a) a metaphor of room layout (b) spatial adjacency

Fig. 5. Convert a metaphor of room layout into spatial adjacency relations indicated by connection lines, drawings by Cui-Xia Ma © Cui-Xia Ma.

represent the temporal and spatial relations among sketches that may implicitly represent procedures such as "plan a visiting path" (the temporal relations in Figure 4, left) or "navigate to a room in a house" (the spatial relations in Figure 5).

The sketch graph provides the user with a simple and easy-to-use tool for visualizing and editing an organization structure of video clips. When a user has a rough idea about the organization, she/he begins to draw some simple sketches for drafting the intent. Based on the simple sketches drawn by the user, the context-aware recommendation technique (to be presented in Section 3.3) recommends to the user the best matched sketch annotations in the database that represent the related video clips. The user can select sketch annotations as nodes, drag and move them to change their positions in the sketch graph, and add or delete sketching connections between nodes, all based on gesture operations. The detailed system implementation is presented in Section 3.4.

3.3. Context-Aware Sketch Recommendation

To offer flexible access to retrieving declarative knowledge in the user's personalized sketch annotations, sketch-based retrieval is useful to facilitate the user's interactive organization of video clips. Most state-of-the-art sketch-based retrieval methods (e.g., [Cao et al. 2011; Sun et al. 2013]) only support ranking and retrieval of images by a single sketch. In our application scenario, however, the semantic meaning of a sketch annotation in a video clip is usually subject to the environmental context of that video clip. For example, the same sketch of a table has different meanings in different video clips, such as a dinner table in a kitchen video and a workbench in a factory video. Therefore, a new context-aware method is needed.

In our study, the definition of context consists of two parts: an environmental context and a relational sketching context.

—The environmental context of a sketch annotation s in a video clip is the set of sketch annotations other than s in the same video clip. For example, a sketch of a refrigerator offers an environmental context of a kitchen video for the sketch of a table.
—Assume that a user is drawing a sketch node in an (uncompleted) sketch graph. A relational sketching context is defined using the uncompleted sketch graph, which consists of the present state (represented by the sketch node being drawn by the user) and the previous states (represented by the sketch nodes already existing in

Table I. Notations Used in Section 3.3

Notation	Meaning	Notation	Meaning
v_i	a video clip	$V = \{v_1, v_2, \ldots, v_n\}$	the set of video clips in database
$T(v_i)$	time duration of v_i	$G(t)$	sketch graph at time t
s_{ij}	sketch annotation in $v_j \in S(V)$	$s(t)$	sketch node at time t
$T(s_{ij})$	time duration of the shot represented by s_{ij}	$P(s_{ij}, v_j \mid s(t), G(t))$	possibility of context-aware recommendation of s_{ij}
$C_b(r)$	a disk of radius r at pixel b	$S(v_i) = \{s_{1i}, s_{2i}, \ldots, s_{n_i i}\}$	the set of sketch annotations in v_i
$H_b(r)$	a histogram in $C_b(r)$ at b	$S(V) = \{S(v_1), \ldots, S(v_n)\}$	all sketch annotations in database
$E_b(r)$	the entropy of $H_b(r)$	pt_i	a part (a 20-vector) in a cluster
PT	the set of all parts in $S(V)$	$s(t) = \{pt_1^1, pt_2^1, \ldots, pt_{n_t}^1\}$	all the parts in $s(t)$
$PT(G(t))$	all the parts in $G(t)$	$s_{ij} = \{pt_1^2, pt_2^2, \ldots, pt_{n_{s_{ij}}}^2\}$	all the parts in s_{ij}
BG	a bipartite graph	$W = \{w_1, w_2, \ldots, w_r\}$	a vocabulary of visual words w_i

the sketch graph). For example, given existing sketches of a wrench and pliers, a user who is drawing a table may have a large possibility of finding a workbench in a factory video.

Next we propose a context-aware sketch recommendation method that takes into consideration the environmental contexts in video clips and the relational sketching contexts in a sketch graph. Table I summarizes all the notations used in describing this method.

Denote the set of video clips in a database by $V = \{v_1, v_2, \ldots, v_n\}$, the sketch annotations in each video clip v_i by $S(v_i) = \{s_{1i}, s_{2i}, \ldots, s_{n_i i}\}$, the currently unfinished sketch graph (at time t) drawn by the user so far by $G(t)$, and the current (possibly unfinished) sketch node drawn by the user by $s(t)$. Sketch annotations similar to $s(t)$ are recommended to the user from $S(V) = \{S(v_1), S(v_2), \ldots, S(v_n)\}$.

Feature extraction in sketch s_{ij}. A sketch is a black-and-white binary image. At each black pixel $b \in s_{ij}$, a histogram $H_b(r)$ is formed in a circular region $C_b(r)$ of radius r. The circle $C_b(r)$ is partitioned into 20 circular bins and each bin k in $H_b(r)$ records the number of black pixels fallen into that bin. We choose such a circular histogram because sketches are usually inaccurate and repeated sketches of the same shape frequently have local distortions like angular squeezing or stretching, thus only using the radial information makes the feature representation insensitive to such angular variations. This has been demonstrated useful in sketch retrieval [Liu et al. 2013b]. The image features may have large variability in scale. We use Kadir and Brady's method [Kadir and Brady 2001] to compute the entropy $E_b(r)$ of $H_b(r)$. By increasing the radius r in $C_b(r)$, the scale for b is determined at the local maxima of $E_b(r)$. For each black pixel b with a scale r and a histogram $H_b(r)$, we further use Kadir and Brady's [2001] method to compute a saliency for b with appropriate scale normalization. Finally, we choose $N = 100$ circular regions $C_b(r)$ with the highest saliency as the features for that sketch.

Shape representation in sketch s_{ij}. Motivated by the success of part-based image categorization [Fei-Fei et al. 2006; Zhao et al. 2011], we represent each sketch s_{ij} using a small set of parts, where each part is a local cluster of some features in the sketch. Given $N = 100$ features in the sketch, we use affinity propagation [Frey and Dueck 2007] to classify the features into optimal clusters in which the number of clusters is automatically and optimally determined. Each cluster represents a part pt_i expressed as a 20-vector $pt_i = \frac{1}{m}\sum_{k=1}^{m} H_{b_k}(r_k)$, where m is the number of features (at pixels b_k, $k = 1, 2, \ldots, m$) fallen into that cluster. For any two parts pt_i and pt_j, their similarity is given by $pt_i \cdot (pt_j)^T$. Usually sketches of different complexities can have parts varying from 2 to 10; some examples are illustrated in Figure 6.

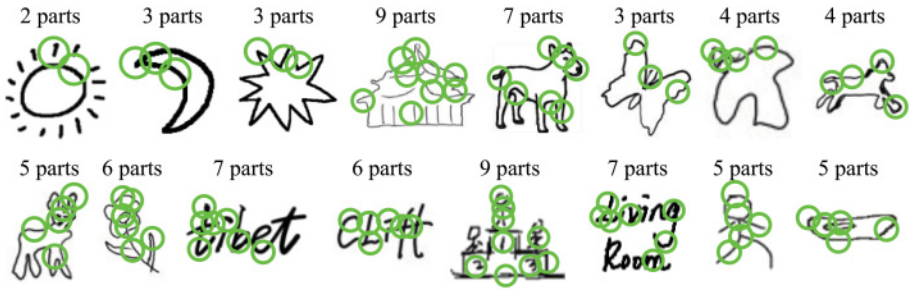


Fig. 6. Part representation of sketches shown in Figure 2(a).

Context-aware recommendation. At an artificial time t, a user is drawing a sketch node $s(t)$ in a sketch graph $G(t)$. We use the following context-aware matching scheme to recommend some most related sketches s_{ij} in $S(V)$ for $s(t)$ such that the shape parts in s_{ij} match shape parts in $s(t)$ and the video clip v_j containing s_{ij} matches the gist of $G(t)$. To do this, we rank all the sketches s_{ij} in $S(V)$ by assigning a possibility $P(s_{ij}, v_j|s(t), G(t))$, which means that, if a user sketches a graph $G(t)$ and is currently drawing a sketch node $s(t)$, the possibility of the user's desired sketch being s_{ij} in a video clip is v_j. We sort all the sketches s_{ij} in $S(V)$ using $P(s_{ij}, v_j|s(t), G(t))$ in a descending order and the top-k sketches in $S(V)$ are recommended to the user for replacing $s(t)$. Ranking with $P(s_{ij}, v_j|s(t), G(t))$ equals to ranking with $P(s_{ij}, v_j, s(t), G(t))$ and we use Bayes rule $P(s_{ij}, v_j, s(t), G(t)) = P(s(t), G(t)|s_{ij}, v_j)P(s_{ij}|v_j)P(v_j)$, where $P(v_j) = \frac{1}{n}$, for all j in $V = \{v_1, v_2, \ldots, v_n\}$. Denote the time durations of the shot represented by s_{ij} and the video clip v_j by $T(s_{ij})$ and $T(v_j)$, respectively. Then $P(s_{ij}|v_j) = \frac{T(s_{ij})}{T(v_j)}$. To compute $P(s(t), G(t)|s_{ij}, v_j)$, we propose to use a weighted bipartite graph matching that has the following characteristics.

—The weighting scheme takes the context information in $G(t)$ and v_j into account.
—The bipartite graph matching allows a partial matching. So it is particularly suitable for matching an $s(t)$ that is drawn only partially and incompletely.

Each sketch is represented by some parts and let PT be the set of all parts in all sketches $S(V)$. Every part $pt_i \in PT$ does not have the same weight for describing sketches. We apply again the affinity propagation to classify the parts in PT into an optimal set of clusters. Each cluster represents a visual word w_i and all visual words form a vocabulary $W = \{w_1, w_2, \ldots, w_r\}$ in which each word w_i represents all parts fallen into the cluster of that word. Now let $s(t)$ and s_{ij} be represented by its parts $s(t) = \{pt_1^1, pt_2^1, \ldots, pt_{n_t}^1\}$ and $s_{ij} = \{pt_1^2, pt_2^2, \ldots, pt_{n_{s_{ij}}}^2\}$. For each part $pt_r^1 \in s(t)$, we assign a weight h_r^1 based on $G(t)$. Let $PT(G(t))$ be all the parts in $G(t)$. Initially, $h_i^1 = 1$ for all i. Then, for each $pt_i \in PT(G(t))$, if pt_i and pt_r^1 relate to the same word, then h_r^1 is increased by one. This means that, if $G(t)$ has more parts fallen into the same cluster of pt_r^1 in the vocabulary W, then pt_r^1 will have a larger weight. Similarly, we assign a weight h_k^2 based on v_j for each part $pt_k^2 \in s_{ij}$. Due to sketch inaccuracy and possible incompleteness, the part number n_t in $s(t)$ may not equal the part number $n_{s_{ij}}$ in s_{ij}. Without loss of generality, let $n_{s_{ij}} \geq n_t$. We construct a bipartite graph BG with nodes $s(t) \bigcup s_{ij}$, and there is an edge $e = (pt_r^1, pt_k^2)$ between any node $pt_r^1 \in s(t)$ and any node $pt_k^2 \in s_{ij}$, with weight $h_r^1 h_k^2 pt_r^1 \cdot (pt_k^2)^T$. The maximum weight matching in BG can be found in $O(n_{s_{ij}}^3)$. We set $P(s(t), G(t)|s_{ij}, v_j)$ to be the maximum weight in BG.

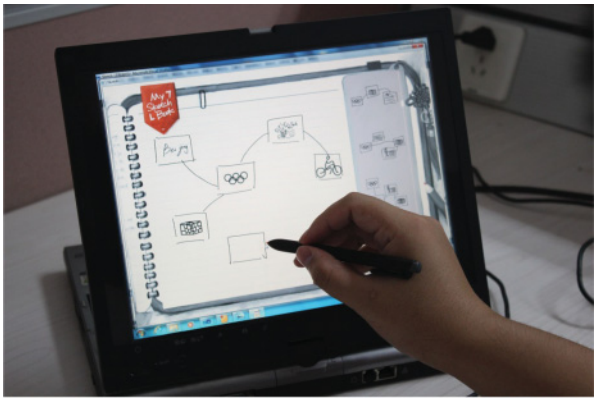


Fig. 7. A prototype system implemented in a display-integrated tablet (TOSHIBA PORTEGE M400).

Scalability. To make our context-aware sketch recommendation useful in a large-scale database with millions of sketches, we propose the following preprocessing and indexing scheme. It was shown in Fei-Fei et al. [2006] that four parts are sufficient to distinguish the type of an image of roughly 300×200 pixels from 101 categories of 9145 images. Their tested image database was later extended to 256 categories of 30607 real images. Accordingly, we choose to use two significant parts from the user's sketch $s(t)$ and use a vocabulary size of 1000 (i.e., 1000 visual words in a vocabulary where each visual word represents a cluster of parts) for a database of millions of sketches. Usually, the sketch graph $G(t)$ has a few to tens of sketch nodes and we use a quantized weight in five levels (20, 40, 60, 80, 100) assigned to the two significant parts of a user's sketch $s(t)$, that is, if the original weight is w, we choose the closest weight in (20, 40, 60, 80, 100) as its quantized weight. Then the number of all the patterns of two parts from 1000 visual words and with weights chosen from quantized weights is $\binom{1000}{2} \times 5^2$. During preprocessing, we match each pattern to the database and build an index structure that stores the top-100 matched sketches. Given that each sketch's ID in the database uses 3 bytes for indexing, the size of the indexing structure is $\binom{1000}{2} \times 5^2 \times 100 \times 3$ *bytes* $\approx$ 3.7GB, which can be preloaded in a common server. Note that each visual word is a 20-vector in which each element is a single-precision floating point taking 4 bytes. Then the file size of the vocabulary is $1000 \times 20 \times 4$ *bytes* $\approx$ 80KB.

During the online recommendation, parts are extracted from the user's sketch $s(t)$ and weighted by the sketch graph $G(t)$. These parts are sorted by weight values and the two parts with largest weights are chosen as significant parts. Their weights are further quantized according to five levels. Given the two significant parts with quantized weights, we look up the index structure and find the top-100 matched sketches in the database. The recommended order of these 100 matched sketches is further updated by using the full set of parts in $s(t)$ with their original (unquantized) weights.

Context-aware sketch recommendation provides a flexible tool for the user's interactive organization of video clips. When a user sketches her/his thought about video content, the most similar annotated sketches in $S(V)$ will be recommended to the user, based on the context information in $G(t)$ and v_j. The user can choose one of recommended sketches, or keep drawing until the desired sketch appears.

3.4. Implementation

The modules of sketch annotation and interactive organization using sketching-graph- and context-aware sketch recommendation are integrated into a consistent and efficient

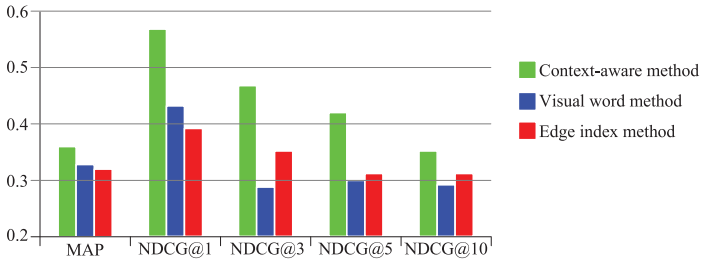


Fig. 8. Accuracy evaluation of three sketch recommendation/retrieval methods: the context-aware method, the edgel index method [Cao et al. 2011], and the visual word method [Ma et al. 2012]. It was observed that the proposed context-aware method had the best accuracy. For details see Section 4.1.

platform with a sketching interface for user-adaptive organization of video clips. We implemented such a prototype system in the *C#* platform and tested it with a display-integrated tablet (Figure 7). Operations on annotations and sketch graph construction with context-aware sketch recommendation are all supported by direct manipulation of gesture commands. Our system focuses on the creation of sketch annotation and narrative structure of the sketch graph. This focus helps to reduce the number of possible interpretations available at the gesture recognition and parsing steps. To support fluid sketching operations, we use Rubine's features [Rubine 1991] to parse single-stroke inputs. Three types of gestures are used in the system:

—basic file operations: New, Save, Clear;
—drawing operations: Ellipse, Rectangle, Loop, Linkage arrows;
—editing operations: Select, Delete, Insert.

To avoid conflicts between making gestural commands and making drawings, a combination of holding and sketching is used to delineate gesture operations from drawing operations.

4. EXPERIMENTS AND USER STUDIES

4.1. Accuracy Evaluation of the Basic Context-Aware Sketch Recommendation Method

We compare the proposed context-aware sketch recommendation with two state-of-the-art sketch-based retrieval methods: the edgel index method [Cao et al. 2011] and the visual word method [Ma et al. 2012]. Five classes of video clips were selected, each class with 4 to 6 video clips. Each class contains a similar object, such as a flag or a car, but the video clips in each class have different environmental contexts, for example, a car in a garage or driving a car on the street. In each video clip, sketch annotations were obtained by auto-extracting line drawings from keyframes followed by the user's simple modification. In total, there were 24 video clips and 452 sketch annotations.

To perform context-aware sketch recommendation, a user was allowed to draw 1 to 2 sketches as the relational sketching context. Then the user drew a sketch and similar video clips were recommended. Three users were invited and each was asked to search three video clips in five classes. Before performing the search test, all the users watched all the video clips and all the sketch annotations. For the edgel index and visual word methods, we assigned the same weights to the context sketches and the searching sketch, while the retrieved videos were ranked by summing up the index values obtained separately from the context sketches and the searching sketch. The performance results including MAP, NDCG@1, 3, 5, 10 are shown in Figure 8, in which the results were averaged over three users and three search performances of each user.

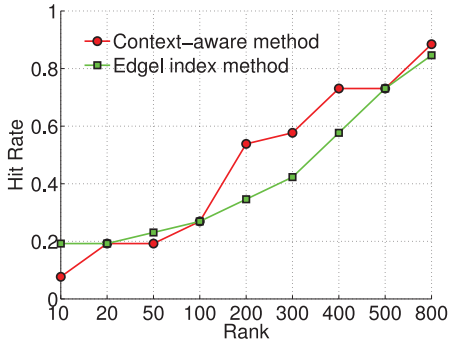


Fig. 9. Scalability performance comparison of our context-aware method with the indexing scheme and the edgel index method in a large-scale database with 1.1 millions of sketches. When the rank K is larger than 100, our context-aware method generally has better search performance than the edgel index method. For details, see Section 4.2.

These results show that our context-aware sketch recommendation achieves the best accuracy.

4.2. Scalability in Specific Video Search Using Context-Aware Sketch Recommendation

In this experiment, the performance of finding a target video clip in a large-scale database using sketches was evaluated for the three methods described in Section 4.1. Twelve students were recruited to manually download video clips from six popular Chinese-language video sharing sites[1], each of which downloaded 50 to 60 video clips per day in 15 days. Ten thousand video clips were downloaded from the Internet and each video clip had a time length of 2 to 5 minutes. The experimental collection consists of 10780 unique videos after removing duplicates. Keyframes were auto-extracted from these video clips and converted into sketches by applying the CLD method [Kang et al. 2007] in the saliency regions. In total, there are 1.1 millions of sketches in the database and each video clip corresponds to 50 to 150 sketches.

To perform search in this large-scale database, six users were invited. Each user randomly chose 10 video clips and watched their sketch annotations. Then each user performed 10 search tasks. For each search task, each user drew 1 to 2 relational context sketches and drew a sketch to search the desired video clip using the three methods detailed in Section 4.1. First, we compare the response time of search using these three methods. For one search task, on a PC (Intel(R) Core(TM) 920 CPU 2.67 GHz, 6GB memory) running Windows 7, the response time of our method with the indexing scheme is 1.32s on average. As a comparison, the response times of the edgel index [Cao et al. 2011] and visual word [Ma et al. 2012] methods are on average 72s and 904s, respectively. These results show that our method is quite suitable for use in an interactive operation and that the visual word method had a very poor scalability. Then we further compare the performance of search results from two scalable methods: our method with the indexing scheme and the edgel index method. We follow Cao et al. [2011] in using "Hit Rate @K" as a measurement. "Hit Rate @K" is the proposition of all the search tasks that rank the target video in the top-k search results averaged over all users. The results are presented in Figure 9, which shows that, when k is larger than 100, our method generally has better performance than the edgel index method.

[1]http://www.youku.com/; http://tv.sohu.com/; http://www.iqiyi.com/; http://v.qq.com/; http://www.letv.com/; http://www.56.com/.

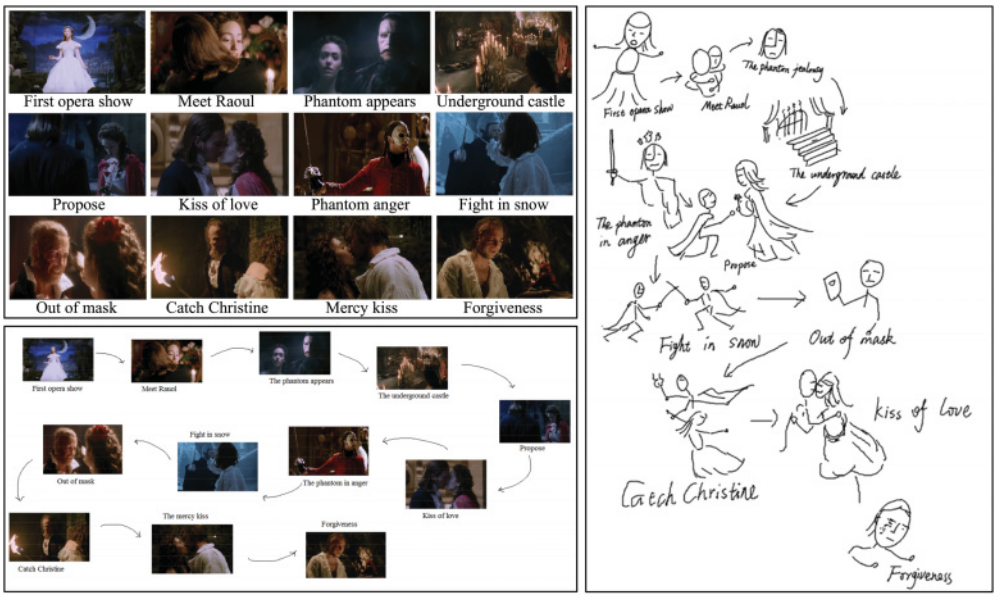


Fig. 10. One example of video organization using three forms by a user. The right drawing by Cui-Xia Ma © Cui-Xia Ma.

4.3. Comparison of Three Visualization Forms for Video Organization

Our proposed sketch graph method uses a visualization form of a sketched storyboard for video organization. Traditionally, keyframes and keywords have been two popular forms for video organization. In the work Ma et al. [2012], the three forms, namely sketches, keyframes, and keywords, are compared and the results showed that sketches have their own advantage. In this study, we further compare three combinatorial forms: (1) sketch graphs in terms of line drawings plus hand-written texts, (2) tiled keyframes with keywords, and (3) keyframes and keywords with sketch connections similar to those in sketch graphs.

Sixteen participants were recruited from the University of Chinese Academy of Sciences, including 9 females and 7 males. Their ages ranged from 20 to 25. They had different majors and varying computer skills. Five different video clips were provided to them, whose lengths ranged from 30s to 500s. The participants were randomly divided into two groups of equal size. One group, the *organizers*, were asked to organize these video clips using three different forms. One example is presented in Figure 10. The other group, the *evaluators*, were asked to rate the resulting video content organization using a variant of the ITU-R 5-point rating labeled as "excellent", "good", "fair", "poor", and "bad", recorded on a numeric scale from 5 to 1. At the end of the user study, an informal interview conducted with the participants regarding their feedback on the different organization forms.

We collected the participants' evaluation results and averaged the scores over five clips. The mean scores are presented in Figure 11. They show that the sketch graphs in terms of line drawings plus hand-written texts have the highest scores. A repeated-measure ANOVA was conducted and the results show that the main effect of different organization forms is significant, $F(2, 14) = 67$, $p < 0.01$.

—There was significant difference between the form (1) (i.e., sketch graphs in terms of line drawings plus hand-written texts, $M = 4.1$, $SD = 0.24$) and the form (2) (i.e., tiled keyframes with keywords, $M = 2.7$, $SD = 0.21$).

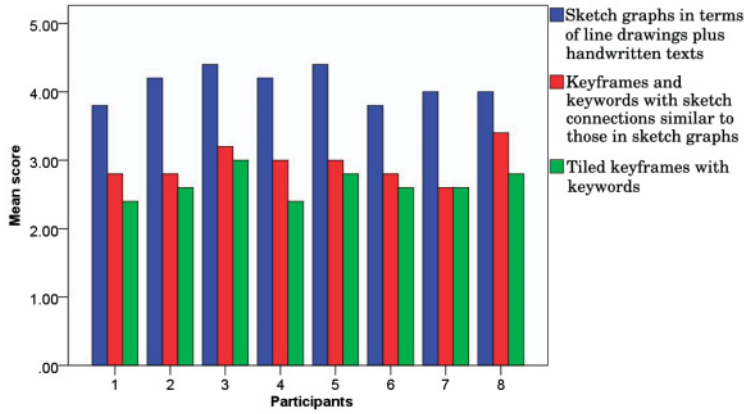


Fig. 11. Mean scores from eight evaluators for three visualization forms, averaged over five different video clips. We observe that sketch graphs are consistently preferred over traditional representations such as keyframes and/or keywords. For details, see Section 4.3.

—There was significant difference between the form (1) (i.e., sketch graphs in terms of line drawings plus hand-written texts, $M = 4.1$, $SD = 0.24$) and the form (3) (i.e., keyframes and keywords with sketch connections similar to those in sketch graphs, $M = 3.0$, $SD = 0.26$), $p < 0.01$).

Based on the results of the informal interviews, 88% of participants (14 of 16) thought that the sketch graph provided a good visualization for video content organization. They thought that the sketches were clear and concise, without interruption from the redundant textural and color information in the keyframes.

4.4. A User Study on Potential Application Scenarios

Three potential application scenarios of the proposed sketch-based approach were studied in this user study and the potential users were targeted as those with amateur interest in media and having a little bit of free time, such as university students or retirees.

—*Scenario 1.* Scenario 1 is where the video maker, the video annotator, and the video organizer are the same person.
—*Scenario 2.* In this scenario, the video maker is one person, while the video annotator and organizer comprise the other.
—*Scenario 3.* In Scenario 3, the video maker, the video annotator, and the video organizer are different people.

Participants. Twenty-one participants from Southwest Jiaotong University took part in the study, including 10 females and 11 males. Their ages were all in the 20 to 30 range. They were from different majors and had different backgrounds: seven were good at computer science, eight were good at mechanical design, and six were good at art painting. According to the questionnaires returned by the participants, they also had different skill levels in video editing and multimedia authoring: six had experience of more than 30 hours and the others had no experience.

Experiment description. A briefing about the questionnaire survey was first presented to all participants. Then a demo video, including two parts, was shown to the participants: Part 1 to show how to annotate video using sketches and Part 2 to illustrate how to use sketch graph and context-aware sketch recommendation for video

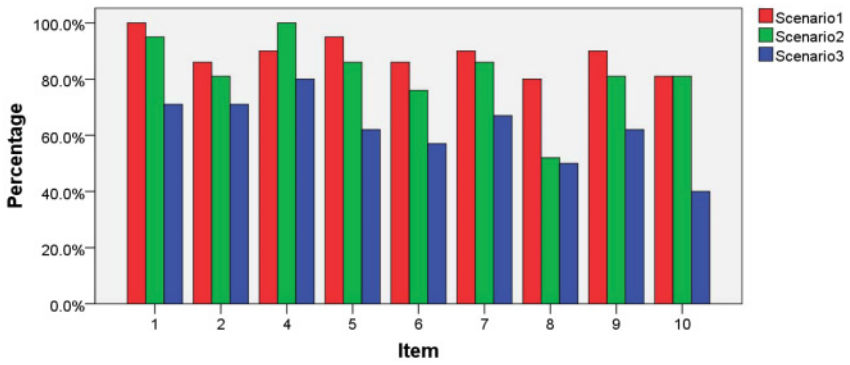


Fig. 12. The percentage of 21 participants who gave "yes" for the nine items in the questionnaire for three scenarios. The results show that the proposed sketch-based video organization is more applicable to Scenarios 1 and 2 than in Scenario 3. For details, see Section 4.4.

content organization. After showing the demo, participants filled in the questionnaire with the three forms corresponding to the three scenarios.

Results. There were nine items in the questionnaire that required judgement by either "yes" or "no" answer. They are: Item 1 (sketch annotations are useful for video organization), Item 2 (sketching interface is appropriate for making annotation), Item 4 (the sketch retrieval results made by context-aware recommendation are useful), Item 5 (sketching interface is appropriate for video organization), Item 6 (understanding narrative structure in sketch graph is easy), Item 7 (sketch graph is useful for video organization), Item 8 (sketch graph is simple for video organization), Item 9 (sketch graph is intuitive for video organization), and Item 10 (sketch graph is efficient for video organization). These items were set to characterize the three features in the sketch-based approach: Items 1 and 2 for sketch annotation, Item 4 for context-aware sketch recommendation, and Items 5 to 10 for sketch graph organization. We sort the order of the three scenarios using the percentage of participants who answered "yes" (Figure 12).

—*Scenario 1.* The nine Items except for Item 4 received the maximum number of "yes".
—*Scenario 2.* The Items 4 and 10 received the maximum number of "yes". Item 10 had the same number of "yes" as in Scenario 1.
—*Scenario 3.* All the nine items received the minimum number of "yes".
—If we consider that the percentage of participants who gave "yes" is a random variable P in the three scenarios, then the results can be summarized as follows. The main effect of the sketch-based approach in the three scenarios was significant, $F(2, 16) = 33.94$, $p < 0.01$. The pairwise comparisons with LSD correction showed that:
  —There was significant difference between Scenario 1 ($M = 88.7\%$, $SD = 6.3\%$) and Scenario 3 ($M = 62.2\%$, $SD = 12.1\%$), $p < 0.01$.
  —There was significant difference between Scenario 2 ($M = 82.0\%$, $SD = 13.5\%$) and Scenario 3 ($M = 62.2\%$, $SD = 12.1\%$), $p < 0.01$.
  —There was no significant difference between Scenario 1 ($M = 88.7\%$, $SD = 6.3\%$) and Scenario 2 ($M = 82.0\%$, $SD = 13.5\%$), $p = 0.082$.

To compare the effect of organization with sketches to those with texts and video frames, there were two items in the questionnaire that needed participants to choose from texts, sketches, and frames. The results were as follows.

—Item 3 (which is better for annotated video contents).
  —In Scenario 1, seven participants (33%) chose texts, 13 participants (62%) chose sketches, and one participant (5%) chose frames.

  - —In Scenario 2, eight participants (38%) chose texts, 10 participants (48%) chose sketches, and three participants (14%) chose frames.
  - —In Scenario 3, fifteen participants (71%) chose texts, two participants (10%) chose sketches, and four participants (19%) chose frames.
- —Item 11 (which is better for visualizing video contents in organization).
  - —In Scenario 1, three participants (14%) chose texts, 12 participants (57%) chose sketches, and six participants (29%) chose frames.
  - —In Scenario 2, four participants (19%) chose texts, 10 participants (48%) chose sketches, and seven participants (33%) chose frames.
  - —In Scenario 3, ten participants (48%) chose texts, five participants (24%) chose sketches, and six participants (28%) chose frames.

We conclude that the proposed sketch-based video organization method is more applicable in Scenarios 1 and 2 than in Scenario 3.

4.5. Efficiency of Sketch-Based Video Organization

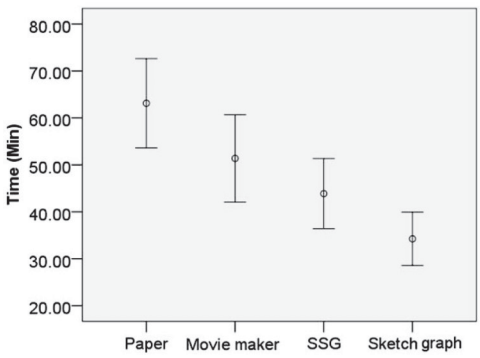
To evaluate the efficiency of the proposed approach using the sketch graph, a user study was conducted for interactive organizations of video clips, with comparison to the following three representative approaches.

- —*Organization with physical paper and pen.* This method fully exploits naturality of users' daily behavior but nothing related to a computer technique.
- —*Organization with the two-layer Scene Structure Graph (SSG) [Ma et al. 2012].* This is a state-of-the-art video authoring and organization method that exploits a natural interface between human sketching behaviors and computer programs.
- —*Organization with Microsoft Movie Maker software.* Movie Maker is a simple and popular software with WIMP interface that has easy and convenient access to video processing and editing.

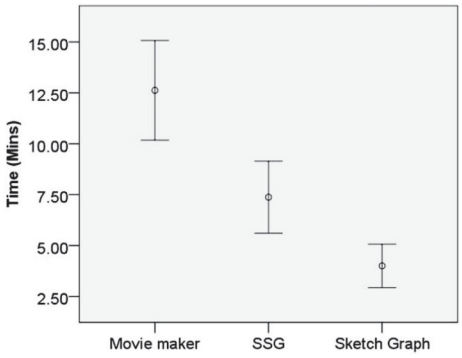
Participants and experiment description. Thirty-two individuals including 20 females and 12 males, aged from 20 to 30, were recruited to participate in the user study. These participants were from the University of Chinese Academy of Sciences, Tsinghua University, and Beihang University, whose majors included computer science, mathematics, biology, and psychology. The participants were randomly divided into four groups, eight participants for each. Each group was required to implement the tasks mentioned shortly using one of the four following methods: (1) pen-and-paper method, (2) two-layer Scene Structure Graph (SSG) [Ma et al. 2012], (3) Movie Maker software, and (4) the proposed sketch-based approach (sketch graph, for short).

Task description. Given a set of 20 video clips whose lengths varied from 30s to 500s, participants were asked to find out the useful ones and organize them into two navigation paths (Tasks 1 and 2). We also asked participants to integrate these two different navigation paths into one by reusing the previous organizations (Task 3). In the testing set of 20 video clips, there were seven video clips showing different places in the Beijing Olympic Sports Center area and five video clips showing different rooms of an apartment. Based on this testing video set, we gave the following three tasks.

- —*Task 1.* This consisted of organizing a tour path inside the Beijing Olympic Sports Center area. For example, the user can design a tour path from the bird nest to the water cube through the National Indoor Stadium.
- —*Task 2.* In this task, the user must organize a visiting path to walk through all the rooms in an apartment. For example, the visiting path can be from livingroom, bedroom, kitchen, to bathroom.
- —*Task 3.* Task 3 is that of integrating the two different paths in Tasks 1 and 2 into one with up to ten sketch nodes in the integration; that is, some sketch nodes in the



(a) performance in video content organization: spent times (mean and standard deviation) of completing Tasks 1, 2, and 3 in four groups; each group had eight participants and used one of the four methods.



(b) performance in reusing existing organization structures: spent times (mean and standard deviation) in Task 3 for three groups; each group had eight participants and used one of three methods.

Fig. 13. Performance comparison of different methods in (a) video content organization and (b) reusing existing organization structures. The results show that the sketch graph method is more efficient than the traditional pen-and-paper method, the method using a mass-market tool Movie Maker, and the SSG sketching method [Ma et al. 2012]. For details, see Section 4.5.

original two sketch graphs had to be deleted. For example, the path can be from the bird nest to the water cube, then to visit the apartment by walking through some rooms, and finally to the National Indoor Stadium.

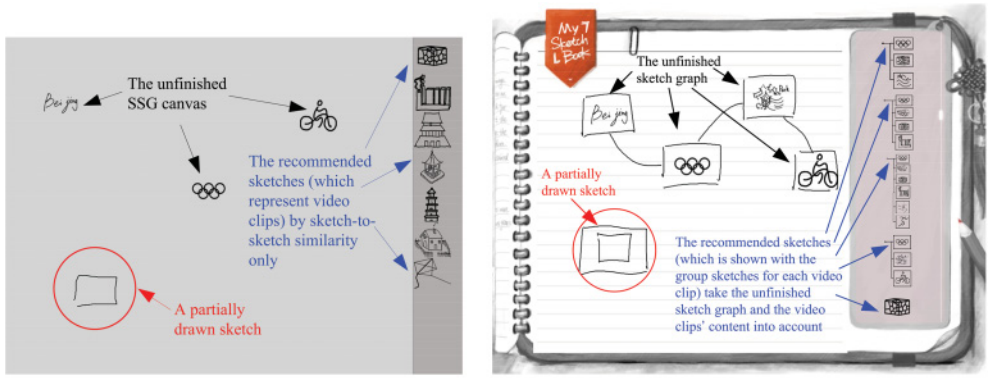
All the tasks were performed in TOSHIBA PORTEGE M400, which is a display-integrated tablet (Intel(R) Core(TM)2 T7200 2.00 GHz) running at Window 7 (Figure 7). One hour's training and practice session for using the four methods was taken with a tutorial before the test.

Experimental results. The spent time (mean and deviation) of completing Tasks 1, 2, and 3 for the four groups using four different methods are summarized in Figure 13(a), which shows that the proposed sketch graph method had the least spent time. A one-way ANOVA method was used and the results showed that the main effect of the different methods was significant, $F(3, 28) = 17.96$, $p < 0.01$.

—There was significant difference between the pen-and-paper method ($M = 63.1$ Mins, $SD = 9.5$) and sketch graph method ($M = 34.3$ Mins, $SD = 5.7$), $p < 0.01$.
—There was significant difference between the Movie Maker ($M = 51.4$ Mins, $SD = 9.3$) and sketch graph method ($M = 34.3$ Mins, $SD = 5.7$), $p < 0.01$.
—There was significant difference between the SSG method [Ma et al. 2012] ($M = 43.9$ Mins, $SD = 7.5$) and sketch graph method ($M = 34.3$ Mins, $SD = 5.7$), $p < 0.05$.

Efficiency of reusing existing organization structures is also important in any organization method. We further made a comparison of spent times in Movie Maker, SSG, and sketch graph methods based on Task 3. The results of spent time using three different methods are summarized in Figure 13(b). The results show that the proposed sketch graph method had the least spent time. A one-way ANOVA test was used and the results showed that the main effect of different methods for changing visiting order was significant, $F(2, 21) = 44.20$, $p < 0.01$. In particular, the results of the pairwise comparisons with LSD correction showed the following.

—There was significant difference between Movie Maker ($M = 12.6$ Mins, $SD = 2.4$) and sketch graph methods ($M = 4.0$ Mins, $SD = 1.1$), $p < 0.01$.



(a) a snapshot of SSG construction (when a user drew a partial sketch on an unfinished SSG canvas, the system recommended several sketches representing video clips by sketch-to-sketch similarity only; the recommended sketches were diverse, some far from the user intent, without taking the unfinished SSG into account)

(b) a snapshot of sketch graph construction (when a user drew a partial sketch on an unfinished sketch graph, the context-aware system recommended sketches that were more accurate than those of SSG recommendation by taking the unfinished sketch graph and the video clips' content into account)

Fig. 14. Difference between (a) sketch retrieval in SSG [Ma et al. 2012] and; (b) context-aware sketch recommendation in the sketch graph.

—There was significant difference between the SSG method ($M = 7.4$ Mins, $SD = 1.8$) and sketch graph method ($M = 4.0$ Mins, $SD = 1.1$), $p < 0.01$.

The experiment results of performance in Tasks 1, 2, and 3 (Figure 13) show that the proposed sketch graph method is more efficient (i.e., has less spent time) than the traditional pen-and-paper method, the classic Movie Maker software, and the SSG method [Ma et al. 2012]. Since both the proposed sketch graph method and the SSG method use sketches for annotation, recommendation, and organization, they are worth comparing so as to make explanations as to why our sketch graph method is better than SSG. There are possibly two reasons.

—SSG uses a two-layer structure: one layer is for visualization and the other is for organization. When a user constructs and changes the order of elements in SSG, she/he has to frequently switch between these two layers. As a comparison, our sketch graph method unifies these two layers into one by designing sketch nodes for declarative knowledge and sketching connections for procedural knowledge. The performance of Task 3, summarized in Figure 13(b), demonstrates the efficiency of the one-layer representation in our sketch graph method.
—The recommendation in SSG relies solely on the sketch similarity. As a comparison, our sketch graph method uses a context-aware sketch recommendation that can better capture the user's intent and is more accurate.

One example demonstrating the inefficiency in SSG sketch recommendation is shown in Figure 14(a): when a user drew a partial sketch on an unfinished SSG canvas, the system recommended a diverse set of sketches (in which some were far from the user intent) without taking the unfinished SSG into account. As a comparison, as shown in Figure 14(b), when a user drew a partial sketch, the context-aware system recommended sketches by taking sketch similarity, the content of the unfinished sketch graph, and the content of video clips into consideration, and thus the recommended results were more accurate than those of SSG recommendation.

5. CONCLUSIONS

In this article a sketch-based method is proposed that allows users to explore quickly and naturally their creative but rough ideas in video organization. The proposed method uses sketch annotations to enrich and extend the semantic knowledge inherent in video clips through modest user interaction. Using annotated sketches as building blocks, users can organize video clips in a sketch graph by searching annotated sketches and combining these sketches in a structural form using different types of connection lines. For facilitating users to efficiently use annotated sketches, a context-aware sketch recommendation technique is suggested and incorporated into the proposed method. All the operations in the proposed method are based on gestures in a sketching interface. Experiments and user studies were performed and the results showed that the proposed sketch-based method offers a promising tool for facilitating users to efficiently organize video clips with an intuitive and natural interaction.

ACKNOWLEDGMENTS

The authors thank the editor and the reviewers for their constructive comments that helped improve this article.

REFERENCES

Brian P. Bailey, Joseph A. Konstan, and John V. Carlis. 2001. DEMAIS: Designing multimedia applications with interactive storyboards. In *Proceedings of the 9th ACM International Conference on Multimedia (MULTIMEDIA'01)*. ACM Press, New York, 241–250.

John B. Best. 1986. *Cognitive Psychology*. West Publishing Company.

Rita Borgo, Min Chen, Ben Daubney, Edward Grundy, Heike Janicke, Gunther Heidemann, Benjamin Hoferlin, Markus Hoferlin, Daniel Weiskopf, and Xianghua Xie. 2011. A survey on video-based graphics and video visualization. In *Proceedings of the Eurographics Conference: State-of-the-Art Reports*. 1–23.

Dick C. A. Bulterman and Lynda Hardman. 2005. Structured multimedia authoring. *ACM Trans. Multimedia Comput. Comm. Appl.* 1, 1, 89–109.

Yang Cao, Changhu Wang, Liqing Zhang, and Lei Zhang. 2011. Edgel index for large-scale sketch-based image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. 761–768.

Carlos D. Correa and Kwan-Liu Ma. 2010. Dynamic video narratives. *ACM Trans. Graph.* 29, 4, 88:1–88:9.

Madirakshi Das and Shih-Ping Liou. 1998. A new hybrid approach to video organization for content-based indexing. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS'98)*. 372–381.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 594–611.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Sci.* 315, 5814, 972–976.

Qiu-Fang Fu, Yong-Jin Liu, Wen-Feng Chen, and Xiao-Lan Fu. 2013. Time course of natural scene categorization in human brain: Simple line-drawings vs. color photographs. *J. Vis.* 13, 9.

Komei Harada, Eiichiro Tanaka, Ryuichi Ogawa, and Yoshinori Hara. 1996. Anecdote: A multimedia storyboarding system with seamless authoring support. In *Proceedings of the 4th ACM International Conference on Multimedia (MULTIMEDIA'96)*. ACM Press, New York, 341–351.

Michal Irani and P. Anandan. 1998. Video indexing based on mosaic representations. *Proc. IEEE* 86, 5, 905–921.

Timor Kadir and Michael Brady. 2001. Scale, saliency and image description. *Int. J. Comput. Vis.* 45, 2, 83–105.

Henry Kang, Seungyong Lee, and Charles Chui. 2007. Coherent line drawing. In *Proceedings of the ACM Symposium on Non-Photorealistic Animation and Rendering (NPAR'07)*. 43–50.

Gaetano Kanizsa. 1979. *Organization in Vision: Essays in Gestalt Perception*. Praeger, New York.

Yong-Jin Liu, Qiu-Fang Fu, Ye Liu, and Xiaolan Fu. 2013a. A distributed computational cognitive model for object recognition. *Sci. China* 56, 9, 1–13.

Yong-Jin Liu, Xi Luo, Ajay Joneja, Cui-Xia Ma, Xiao-Lan Fu, and Da-Wei Song. 2013b. User-adaptive sketch-based 3d cad model retrieval. *IEEE Trans. Autom. Sci. Engin.* 10, 3, 783–795.

Cui-Xia Ma, Yong-Jin Liu, Hong-An Wang, Dong-Xing Teng, and Guo-Zhong Dai. 2012. Sketch-based annotation and visualization in video authoring. *IEEE Trans. Multimedia* 14, 4, 1153–1165.

Cui-Xia Ma, Yong-Jin Liu, Hai-Yan Yang, Dong-Xing Teng, Hong-An Wang, and Guo-Zhong Dai. 2011. KnitSketch: A sketch pad for conceptual design of 2d garment patterns. *IEEE Trans. Autom. Sci. Engin.* 8, 2, 431–437.

Tao Mei and Xian-Sheng Hua. 2008. Structure and event mining in sports video with efficient mosaic. *Multimedia Tools Appl.* 40, 1, 89–110.

Tao Mei, Bo Yang, Shi-Qiang Yang, and Xian-Sheng Hua. 2008. Video collage: Presenting a video sequence using a single image. *Vis. Comput.* 25, 1, 39–51.

Emily Moxley, Tao Mei, and Bangalore S. Manjunath. 2010. Video annotation through search and graph reinforcement mining. *IEEE Trans. Multimedia* 12, 3, 184–193.

Paul A. Rodgers, Graham Green, and Alistair Mcgown. 2000. Using concept sketches to track design progress. *Des. Studies* 21, 5, 451–464.

Dean Rubine. 1991. Specifying gestures by example. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'91).* ACM Press, New York, 329–337.

Xinghai Sun, Changhu Wang, Avneesh Sud, Chao Xu, and Lei Zhang. 2013. MagicBrush: Image search by color sketch. In *Proceedings of the 21st ACM International Conference on Multimedia (MM'13).* ACM Press, New York, 475–476.

Zhenbang Sun, Changhu Wang, Liqing Zhang, and Lei Zhang. 2012. Free hand-drawn sketch segmentation. In *Proceedings of the 12th European Conference on Computer Vision (ECCV'12)*. Lecture Notes in Computer Science, vol. 7572. Springer, 626–639.

Ba Tu Truong and Svetha Venkatesh. 2007. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Comm. Appl.* 3, 1.

Jingdong Wang and Xian-Sheng Hua. 2011a. Interactive image search by color map. *ACM Trans. Intell. Syst. Technol.* 3, 1, 12:1–12:23.

Meng Wang and Xian-Sheng Hua. 2011b. Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell. Syst. Technol.* 2, 2.

Meng Wang, Xian-Sheng Hua, Jinhui Tang, and Hong Richang. 2009. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Trans. Multimedia* 11, 3, 465–476.

Cheng-Chi Yu, Yong-Jin Liu, Matt Tianfu Wu, Kai-Yun Li, and Xiaolan Fu. 2014. A global energy optimization framework for 2.1d sketch extraction from monocular images. *Graph. Models* 76, 507–521.

Jin-Kai Zhang, Cui-Xia Ma, Yong-Jin Liu, Qiu-Fang Fu, and Xiao-Lan Fu. 2013. Collaborative interaction for videos on mobile devices based on sketch gestures. *J. Comput. Sci. Technol.* 28, 5, 810–817.

Yu-Jin Zhang and Haibao Lu. 2002. A hierarchical organization scheme for video data. *Pattern Recogn.* 35, 11, 2381–2387.

Bin Zhao, Li Fei-Fei, and Eric P. Xing. 2011. Large-scale category structure aware image categorization. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS'11).* 1251–1259.

Xingquan Zhu, Ahmed Elmagarmid, Xiangyang Xue, Lide Wu, and Christine Catlin. 2005. Towards hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Trans. Multimedia* 7, 4, 648–666.

Received September 2013; revised April 2014; accepted June 2014