VSumVis: Interactive Visual Understanding and Diagnosis of Video Summarization Model

GUODAO SUN, Zhejiang University of Technology, China
HAO WU, LIN ZHU, CHAOQING XU, HAORAN LIANG, BINWEI XU, and
RONGHUA LIANG, Zhejiang University of Technology

With the rapid development of mobile Internet, the popularity of video capture devices has brought a surge in multimedia video resources. Utilizing machine learning methods combined with well-designed features, we could automatically obtain video summarization to relax video resource consumption and retrieval issues. However, there always exists a gap between the summarization obtained by the model and the ones annotated by users. How to help users understand the difference, provide insights in improving the model, and enhance the trust in the model remains challenging in the current study. To address these challenges, we propose VSumVis under a user-centered design methodology, a visual analysis system with multi-feature examination and multi-level exploration, which could help users explore and analyze video content, as well as the intrinsic relationship that existed in our video summarization model. The system contains multiple coordinated views, i.e., video view, projection view, detail view, and sequential frames view. A multi-level analysis process to integrate video events and frames are presented with clusters and nodes visualization in our system. Temporal patterns concerning the difference between the manual annotation score and the saliency score produced by our model are further investigated and distinguished with sequential frames view. Moreover, we propose a set of rich user interactions that enable an in-depth, multi-faceted analysis of the features in our video summarization model. We conduct case studies and interviews with domain experts to provide anecdotal evidence about the effectiveness of our approach. Quantitative feedback from a user study confirms the usefulness of our visual system for exploring the video summarization model.

CCS Concepts: • **Human-centered computing → Visualization**; **Visualization systems and tools**;

Additional Key Words and Phrases: Visual Analytics, video summarization, video visualization, machine learning, multimedia visual analysis

The work is jointly supported by National Key Research and Development Program of China (2020YFB1707700), National Natural Science Foundation of China (61972356, 62036009), Fundamental Research Funds for the Provincial Universities of Zhejiang (RF-A2020001), Zhejiang Provincial Natural Science Foundation of China (LY19F020026), and Zhejiang Provincial Key Research and Development Program of China (2019C01009).

Authors' addresses: G. Sun, H. Wu, L. Zhu, C. Xu, H. Liang, B. Xu, and R. Liang (corresponding author), Zhejiang University of Technology, No. 288 Liuhe Road, Hangzhou, China; emails: guodao@zjut.edu.cn, {hooscope, zhulin3695, superclearxu}@gmail.com, haoran@zjut.edu.cn, xubinwei521@gmail.com, rhliang@zjut.edu.cn.

2157-6904/2021/06-ART41 $15.00
https://doi.org/10.1145/3458928

ACM Reference format:
Guodao Sun, Hao Wu, Lin Zhu, Chaoqing Xu, Haoran Liang, Binwei Xu, and Ronghua Liang. 2021. VSumVis: Interactive Visual Understanding and Diagnosis of Video Summarization Model. *ACM Trans. Intell. Syst. Technol.* 12, 4, Article 41 (June 2021), 28 pages.
https://doi.org/10.1145/3458928

1 INTRODUCTION

With the rapid development of multimedia network technology and the popularity of video capture devices, especially in the era of mobile Internet, large-scale video sharing websites such as YouTube have sprung up. Multimedia video resources have expanded rapidly, which may lead to issues such as the increasing amount of storage resource consumption and challenges in mass video retrieval. In the face of such dilemmas, video summarization is often seen as an effective way to promote this problem [1].

The video summary techniques generate a concise summarization in an automated or semi-automatic manner by extracting meaningful segments or keyframes from original videos [2]. The video summary could preserve the diversity of the original video while reducing storage space and transmission time, which helps users quickly browse or retrieve videos. Early video summary techniques are mainly based on the manual sampling of keyframes in the video. Although such methods are easy to implement, they do not consider the coherence of video content. Following research relies on video and image processing technology, which utilizes features such as color, texture, or shape to conduct motion detection, keyframes detection, or scene clustering [3].

In the latest video summarization study, researchers usually combine typical machine learning methods with artificially designed features such as motion and aesthetics to identify behaviors in the video to obtain more accurate video summarization [4]. However, these methods face many challenges in practical applications. First, machine learning–based methods sometimes result in a poor quality of video summarization due to inappropriate feature extraction and model selection. For example, too few features may lead to low model accuracy, and too many features may result in an over-fitting phenomenon. In addition, due to the lack of datasets, some features are not considered during the training process. Therefore, the lack of intuitive and effective tools to analyze the relationship between model results and input features is a daunting challenge. Second, difference or temporal delays with respect to the determination of significant frames annotated by humans and predicted by the model are commonly found in current experiments. However, limited by the uncertainty of the parameters in the time series, it is difficult to compare these differences with traditional calculation methods. Thus, how to help users analyze the difference between the manual annotation score and the saliency score predicted by the model is the second challenge we face. Third, the video summarization result is closely related to video types. For example, it is difficult to apply one model to extract the summary of static, moving, or egocentric videos simultaneously. Therefore, assisting researchers to explore video content and analyze the adaptability of the model remains a major challenge.

In this study, we propose a visual analysis system named VSumVis with multi-video feature examination and multi-level exploration (i.e., video, event, and frame), which could help users explore and analyze video content, as well as the intrinsic relationship that existed in our video summarization model. Through an iterative design study with domain experts, we collect a set of requirements for understanding the feature-based video summarization model and exploring video content, then identify visual design goals for constructing a visualization view. We further propose a novel visualization system to support interactive and insightful analysis into the feature-based video summarization model at three levels of detail. In addition, we propose a rule-based sampling algorithm to relax the visual occlusion problem caused by nodes overlapping. We then perform case studies with domain experts on the SumMe dataset, which confirm the effectiveness of our system. Finally, we conduct a quantitative user study to validate the usefulness of our visualization system for analyzing video content and the intrinsic relationship that existed in our video summarization model. The key contributions of this work are as follows.

- We present a comprehensive analysis of multi-feature pattern extraction and multi-level exploration for our video summarization model.
- We propose a visual analysis system that allows interactive exploration of the intrinsic relationship among video summarization models, video features, and original frames.
- We provide case studies based on the SumMe dataset to evaluate the effectiveness of our methods and a user study that verifies the usefulness of our visual analysis system.

2 BACKGROUND AND RELATED WORK

This section reviews related work on video summarization techniques, multimedia visual analysis, and explainable machine learning.

2.1 Video Summarization Techniques

Video summarization is a brief overview of video content, which is an important part of video retrieval technology. It reserves key content and consistency of the original video while compressing the raw video, which provides users an intuitive way to view the video non-linearly. Video summarization techniques mainly have two categories [5]. One is static video summarization, which generates a series of keyframe objects composed of corresponding semantic units to summarize the lens content and supports video fast navigation. The other one is the dynamic video summarization, which maintains the intrinsic features of the video whose content dynamically changes over time. There exists a clear difference between static and dynamic video summarization. Static video summarization only considers the keyframe objects, ignores audio information, and generates faster than the dynamic one. Dynamic video summarization is richer than the static one and combines information such as images, sounds, and text. In this work, we target the explanation of the dynamic video summarization model, which relies on machine learning techniques. Our work aims to visually explore the difference between the video summarization results obtained by the machine learning model and human annotation, as well as the cause.

In recent years, due to the rapid development of artificial intelligence, automatically generated dynamic video summaries through machine learning has become a hot topic in the video summarization field. Gygli et al. [6] predicted the interest of video frames by linear regression and selected the keyframes with the highest interest score. Zhang et al. [7] proposed a novel subset selection technique that uses artificially created summaries to automate keyframe acquisition. Zhou et al. [8] proposed an unsupervised video summarization method based on deep reinforcement learning. At the same time, video summarization also provides a number of new ways to view videos. A typical method is to use a collection of pictures or a storyboard using keyframes [9]. For example, Rother et al. [10] propose a digital tapestry that automatically synthesizes an output image. The other one is a dynamic representation such as video skimming [11] and video redirection method [12]. Based on machine learning technology, we deeply examine the video summarization model proposed by us and present a new novel visual analysis system to perform in-depth multi-featured pattern extraction and multi-level exploration for the video summarization model.

2.2 Multimedia Visual Analysis

Multimedia data includes audio, images, and video, and research has been progressing in a single modal analysis. In the field of multimedia, data is presented as a history of progress in witnessing historical development. The emergence of sound cards in the 1980s marked the computer's ability to process audio. As digital images became the archives of the general public, research has evolved from the traditional edge of individual images or frames to the study of video content. For example, Aigrain et al. [3] studied images based on color, texture, shape, and spatial similarity. The Internet tide has brought us massive multimedia information, and the development of technologies such as machine learning and image processing has spawned a hot spot in video content research. Smeulders et al. [13] focused on image processing, pattern analysis, and machine learning. In addition to discussing the basic features such as color and texture, they also examined certain more advanced topics and cover machine learning topics such as similar matching and semantic interpretation, as well as database topics such as image indexing, storage, and querying. However, most of these analyses of multimedia data tend to stay in the research of improving the technology itself rather than evaluating and explaining them. Visual analysis provides a new perspective for multimedia analysis. The visual analysis transforms information into a computable form and provides dynamic analysis intuitive visual guidance through human-computer interaction and various forms of a visual interface [14]. The combination of visualization and multimedia analysis has greatly contributed to the development of multimedia visual analysis.

Many visual analysis systems [15, 16] have been proposed to aid in multimedia analysis and knowledge discovery. The proposed systems in related literature could be divided into two categories, i.e., analysis of video content and presentation of raw video. Regarding the first one, Wu and Qu [17] use computer vision technology to capture multimodal content in video to explore language and non-verbal skills in TED presentations. Zeng et al. [18] proposed a visual analysis system to explore the consistency of multimodal emotions in speech videos. Kurzhals et al. [19] describe the film content in multiple layers of descriptive features and apply it to content analysis of popular movies. Liao et al. [20] proposed a visual-based active label method for video labeling. Regarding the second one, Liu et al. [21] proposed a nonlinear dynamic representation called a spiral band that summarizes the video in a smooth spiral pattern. Mei et al. [22] provided a compact composite tile for a video sequence in several different layouts. Similar schematic diagrams [9, 23, 24] and comic book–like graphics layouts [25] are also proposed for displaying video summarization. In addition, multi-level granularity analysis presents challenges for multimedia visual analysis. In the structure perspective, video is divided into short clips and frame levels, while in the content perspective, it is divided into different granularity-like events, sequence, or lens. Certain systems expand video content in a temporal sequence and analyze it at the video level in a linear perspective [26, 27]. Other systems analyze videos at a single level of granularity, e.g., Renoust et al. [28] utilized graphic design to visualize videos frame by frame. To solve these problems, our proposed visual analysis system could support users to conduct multi-feature pattern extraction and multi-level exploration. Assisted with other coordinated views, users could obtain a more detailed level analysis.

2.3 Explainable Machine Learning

In recent years, machine learning has been developed rapidly and has achieved remarkable success in many research fields, e.g., artificial intelligence. Machine learning applications have an extremely profound influence on our society. However, due to the inherent black box of machine learning models [29], it is almost impossible for humans to understand the machine learning mechanisms and it lacks interpretation of how a machine model comes up with a specific decision. Researchers have been enthusiastic about the decision-making progress and the expansibility of a machine learning model. The visualization community has developed a variety of visual analytic methods helping domain experts open black boxes of machine learning models [30–33]. Up to the present, plenty of visual analysis methods have been proposed to help researchers with the interpretability and performance diagnosis of a model. Zhao et al. [34] explored the random forest mechanism by designing multiple coordinated linkage views. Das et al. [35] proposed a novel visualization technique that allows a user to detect multiple machine learning models and guides users to build machine learning models that fit specific domain tasks. Similarly, visual analysis works also contribute to the interpretability of neural networks [36–41], which reveals the inner logic of prediction processes. Some studies focus on dealing with the data input to machine learning models. Xiang et al. [42] developed a scalable data correction algorithm that aims at efficiently processing thousands of tagged data. Kande et al. [43] proposed a visual analysis tool for assessing quality issues in tabular data. It helps analysts discover and analyze anomalies hidden in the dataset. Liu et al. [44] developed a visual analysis tool that helps experts validate undefined tags and unreliable labeling in crowdsourcing. When it comes to the visualization of input-output of machine learning models, Ming et al. [45] extracted standardized representations on rule-based knowledge from the input-output pattern of the classifier. Different from those visual analysis works on machine learning model interpretation and data explanation. We offered a visual analysis system targeting comparing the machine learning results and the manual annotation results in video summaries. By using multiple coordinated linkage views, users are allowed to interactively explore the intrinsic relations among machine learning models, video features, and raw content.

3 DESIGN PROCESS AND REQUIREMENT ANALYSIS

Following a user-centered design methodology, we further investigated essential literature closely related to the topic about video analysis [46], temporal analysis [47], and model analysis [48–50]. In this section, we first describe how we identified and updated research questions with domain experts through an interview, and summarize the distilled requirements.

3.1 Design Process

VSumVis aims to help users explore the video summarization model and analyze the intrinsic relationship that existed in model input, output, and video content. To understand the current practice and challenges of the feature-based video summarization model, our design process started with a technology-sharing meeting. During the meeting, we conducted contextual inquiries to collect information about the video summarization and difficulties encountered by model understanding, which motivated the initial design of our system. In the later stages, we adopted an iterative development approach by carrying out bi-weekly meetings with a group of domain experts comprising one assistant professor who engaged in computer vision area research for more than 9 years and four postgraduate students. During the meetings, we collected experts feedback on our early prototypes and updated the system design. Specifically, the experts play a significant role in proposing the feature-based video summarization model for this work, and their expertise in computer vision helps us gain a deeper understanding of the model. Their experience makes them better aware of the difficulties that users may encounter in improving their model analysis. At the same time, they have deep insights into the limitations of traditional methods for model analysis. Thus, they have a great interest to discuss the kinds of requirements in a feature-based model.

3.2 Requirement Analysis

The requirements are formulated throughout the 3 months and we summarize them as follows:

R1 What perspectives could be employed to analyze the model? How to associate the input and output of the model with the corresponding video content. How to discover the patterns among them in a more intuitive way.

R2 There are many diverse types of videos in the video collections. How could we help users quickly locate videos of interest for effective analysis and exploration?

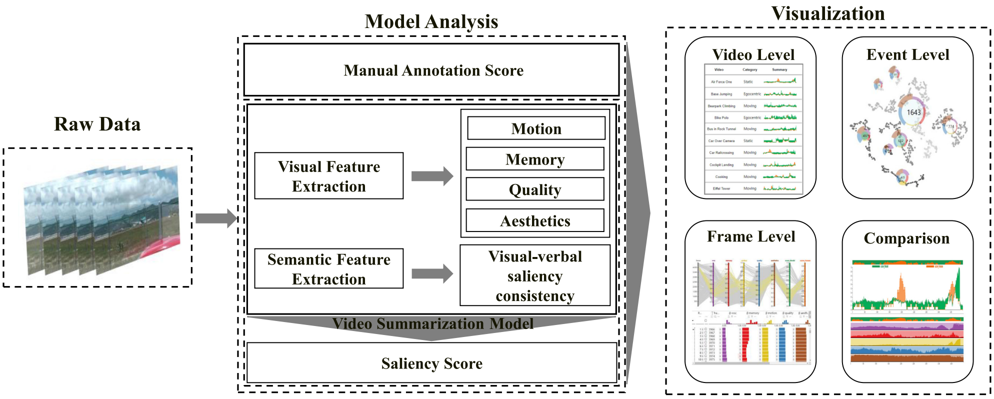


Fig. 1. The pipeline of our work is divided into three steps. First, feature extraction is performed on the SumMe video collections, and the features are further reconstructed and fed into our machine learning model. The output of the model mainly consists of the saliency score of each image, and a descriptive score of five predefined elaborated features. Second, manual annotation score, saliency scores predicted by the model, and scores obtained by the five descriptive features are further visually explored and compared at three levels, namely, video level, event level, and frame level, along with a temporal view. Finally, users could interact with the system to diagnose the model and provide explanations and feedback for the model design.

R3 How are the scores of manual annotation and the scores predicted by the model distributed over time? What is the difference between the two scores on each frame? When will significant anomalies occur?

R4 How are the input and output related in the feature-based video summarization model? What is the dynamic correlation between the feature scores extracted from video and the scores predicted by the model?

R5 How do the extracted features reflect the content in the raw video? Is there any temporal delay compared with the manual annotation along with the sequential frames?

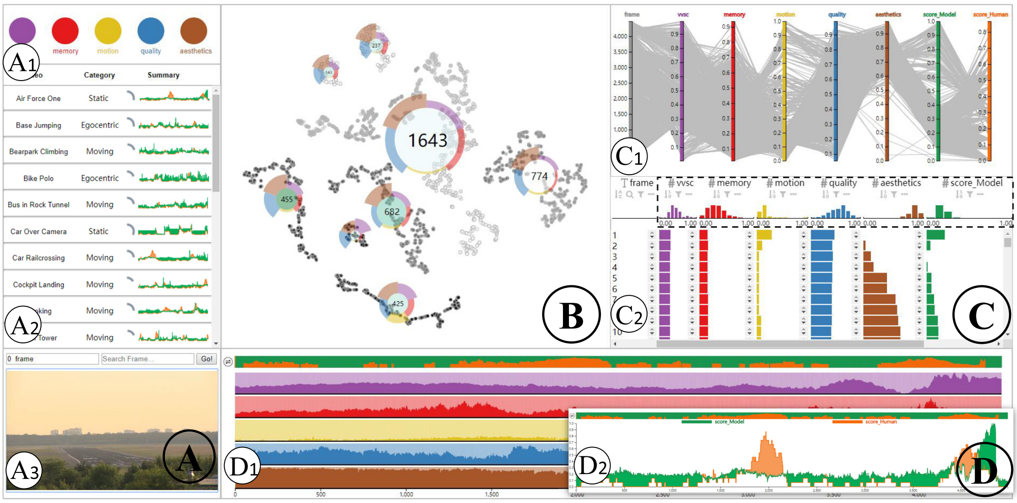
R6 Is there a certain similarity and connection among the frames in the video, and how does one find out the connection between them and define this connection? How does one present it intuitively through visualization?

These requirements help us draw the roadmap for our visual system and shape the detailed principles of visual design.

4 SYSTEM OVERVIEW

Guided by the aforementioned requirements, we designed and developed VSumVis, a visual analytics system for interactively exploring a feature-based video summarization model. In this section, we first describe our analysis pipeline and then briefly introduce each view of the system.

Our pipeline (see Figure 1) mainly consists of two parts: model analysis and interactive visualization. The model analysis part is to extract elaborated visual features and semantic features for each frame in the SumMe dataset video. The features are fed into our machine learning model to extract the corresponding saliency score for each frame. The process is described in detail in Section 4. In the interactive visualization phase, users could examine the patterns behind the model through three-level exploration, namely, video, events, and frames level. Different views are closely coordinated, and friendly and intuitive interactions are equipped within the system for exploration among descriptive features, manual annotation results, and model output.



eo	Category	Summary
Air Force One	Static	
Base Jumping	Egocentric	
Bearpark Climbing	Moving	
Bike Polo	Egocentric	
Bus in Rock Tunnel	Moving	
Car Over Camera	Static	
Car Railcrossing	Moving	
Cockpit Landing	Moving	
king	Moving	
Tower	Moving	

0 frame | Search Frame... | Go!

Fig. 2. Our visualization system supports visual exploration within video, event, and frame levels. (A) The video view provides users with overall distribution of summarization of video collections, which allows them to quickly browse and locate videos of interest. (B) The projection view presents a multi-level analysis process to integrate video events and frames into a single view with clusters and nodes visualization. (C) The detailed view fully displays the detailed feature information of each frame as well as their relationship. (D) The sequential frames view combines manual annotation score, saliency score predicted by the model, and descriptive feature score into a pixel bar chart and Playfair-style chart to analyze the difference as well as potential temporal delay.

Our system has four main views (see Figure 2). The video view (see Figure 2(A)) shows a list of videos outlining the differences between saliency score by the model and manual annotation score. Users could select videos of interest according to their preferences for further exploration. A video panel is presented at the bottom of the video view to help users directly view the raw video content. The projection view (Figure 2(B)) presents the clustering result of all frames in a single video. This view embeds the characteristics of each frame including the scores manually annotated by humans and the saliency score produced by our model. The lineup view and parallel coordinate view (Figure 2(C)) display the relationship among the scores of different descriptive features as well as the score from model and human. The sequential frames view (Figure 2(D)) provides a straightforward visual comparison among the above numerical scores over video frames, which allows users to examine significant frames in the project and detailed view. Section 5.3 of this article will explain Figure 3 in detail.

5 MODEL ANALYSIS

This section elaborates on the details of model analysis, which contains three main parts, i.e., data description, descriptive feature extraction, and our saliency generation model.

5.1 Data Description

The raw dataset, known as SumMe [6] in the computer vision research field, is composed of 25 videos, each of which lasts about 1–6 minutes. According to the video capture devices, all the videos are divided into three categories: static video, egocentric video, and moving video. The static videos were captured by fixed cameras, the egocentric videos were recorded by head-mounted video devices, and the moving videos were taken while the camerists were moving

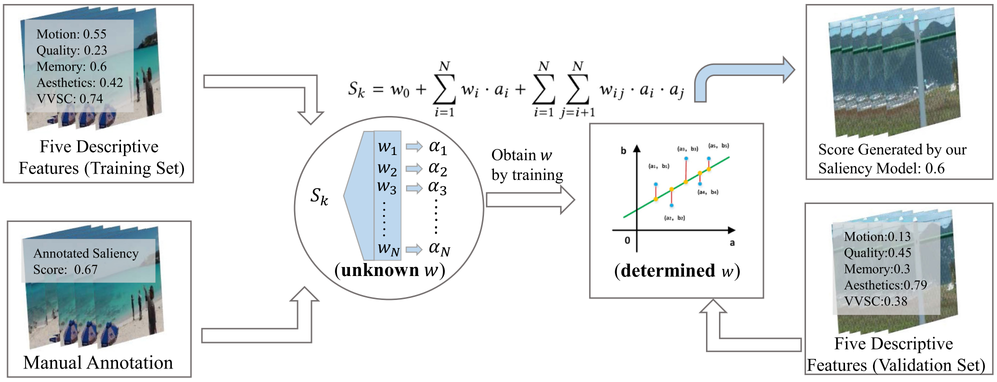

Motion: 0.55
Quality: 0.23
Memory: 0.6
Aesthetics: 0.42
VVSC: 0.74
Five Descriptive Features (Training Set)
Annotated Saliency Score: 0.67
Manual Annotation
$S_k = w_0 + \sum_{i=1}^{N} w_i \cdot a_i + \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_{ij} \cdot a_i \cdot a_j$
S_k
$w_1 \Rightarrow \alpha_1$
$w_2 \Rightarrow \alpha_2$
$w_3 \Rightarrow \alpha_3$
$w_N \Rightarrow \alpha_N$
(unknown w)
Obtain w by training
b
a
0
(a_1, b_1)
(a_2, b_2)
(a_3, b_3)
(a_4, b_4)
(a_5, b_5)
(determined w)
Score Generated by our Saliency Model: 0.6
Motion:0.13
Quality:0.45
Memory:0.3
Aesthetics:0.79
VVSC:0.38
Five Descriptive Features (Validation Set)

Fig. 3. Our saliency score generation model. Each image frame is manually annotated and put into the model together with the five descriptive features. Through repeated iterations, the weights in Formula (1) are obtained and used for predicting the saliency score of all frames in other videos.

forward. Video annotation is further applied to all the videos in the SumMe dataset with each video annotated by 15–18 people. These people are required to make a video summarization for each video. After the manual annotation, each frame will obtain an importance score that indicates how significant the frame is from the average perspective of participating users.

5.2 Descriptive Feature Extraction

To help researchers obtain a deeper understanding of each frame, and improve the summarization accuracy in the videos, co-authors from the computer vision research field of this article performed a series of descriptive feature extractions. The features we extracted fall into two categories: visual feature and semantic feature.

5.2.1 Visual Features. The visual feature indicates the factors that play a vital role in attracting video viewers, such as colors, image quality, motion cues, and visual saliency. It allows users to intuitively perceive the video content and provides a visual perspective for users to analyze the video summarization model (R1 and R5). Through the examination of current videos and literature review, four features have been extracted from each frame in the videos, namely, *motion* feature, *memory* feature, *quality* feature, and *aesthetics* feature.

Motion Feature. Generally, humans pay more attention to moving objects than stationary objects and may be more interested in the motion of the target object. Therefore, to reduce the motion interference of other objects in the background. We use visual saliency [51] to predict the area of people's visual interest, and then extract the *motion* feature of the target object [52].

Memory Feature. Impressive fragments or anti-forgetting information would be kept in one's mind after watching a movie or a short video. We regard those video fragments as important components that should be included in video summarization. Based on current works [53] and [54], we use a memory dataset *LaMem* to train a deep neural network to predict *memory* scores.

Quality Feature. Quality assessment is used to assess whether the image is noisy, fuzzy, compressed, or in other conditions. A low-quality video could be both visually and audibly discomforting. To obtain an excellent quality of the video summary, we introduce the quality assessment prediction proposed by Hossein Talebi [55]. It employs a deep convolutional neural network to predict the distribution of people's opinion scores, and the obtained results have a high correlation with human perception. We use the quality-related data of the image to train MobileNet, and its output is the *quality* score we use in our current work.

Aesthetics Feature. A video's lack of aesthetics may bring a poor viewing experience. The aesthetic assessment is to quantify the aesthetic features in the image, which helps generate a summary of the video that people may be interested in. Our work employs the aesthetic prediction proposed in [55] to enhance our model's ability to recognize the beauty of the picture. The output value obtained from the MobileNet based on image-related data lies between 0 and 1, which is considered as *aesthetics* score.

5.2.2 Semantic Feature. Video contains a wealth of semantic information, which is not negligible in extracting key content from the raw video. It is also the key judgment basis for measuring the importance of frames. Therefore, it is difficult to merely rely on the above visual features to simulate a multi-dimensional "taste" of human vision. Thus, it is vital to combine semantic features with the above visual features to obtain a more comprehensive understanding of video summarization. Introducing semantic features into the video summarization model could better help users discover a video of interest in the video collections (R2).

Visual-verbal saliency consistency Feature: Verbal saliency [56] indicates the meaningful words digging out from a series of sentences describing a specific image. It shows a high correlation between visual perception and semantic description of an image. The consistency of an image description would be used as a critical indicator of semantic significance. Therefore, we introduce visual semantic saliency as identification to pick out the frames that are highly correlated with the corresponding video theme. The abbreviation of **visual-verbal saliency consistency** is ***VVSC***.

5.3 Saliency Generation Model

Based on the above visual and semantic features, we need a mathematical model to incorporate them to generate a quantitative score for later analysis (R3). In this section, we introduce our saliency generation model to efficiently obtain saliency scores for each frame in the videos. As shown in Figure 3, based on the feature extraction, we could obtain the above five descriptive feature scores for each frame. The five feature scores are normalized and further fed into our model to generate a saliency score (i.e., S_k) for each frame. S_k is defined as follows:

$$S_k = w_0 + \sum_{i=1}^{N} w_i \cdot a_i + \sum_{i=1}^{N} \sum_{j=i+1}^{N} w_{ij} \cdot a_i \cdot a_j, \tag{1}$$

where a_i is the score of each normalized feature, N is the number of features, and w_0, w_i, and w_{ij} are the weights which are estimated using least squares. The obtained S_k will be further compared frame by frame with the scores annotated by users.

After obtaining the scores of all frames, the original video is divided into multiple fragments. We further calculate the average importance score of each segment, sort them in descending order, and select segments that match our requirements (e.g., segment length is less than or equal to 15% of the video length). As shown in the right red box of Figure 4, the selected video segments are highlighted in the sequential frame view. These segments are then chronologically organized to form the final video summary.

6 VISUAL DESIGN

This section introduces our visual design. Figure 2 shows our user interface that has four main views. The video view (Figure 2(A)), which displays summary information for all the videos in the video collections, helps users quickly locate the video of interest and explore. The projection view (Figure 2(B)), which shows a two-dimensional distribution of all frames in a video, is presented to explore event-level and frame-level information. The detailed view (Figure 2(C)) shows the details of descriptive feature scores, manual annotation score, and saliency score produced

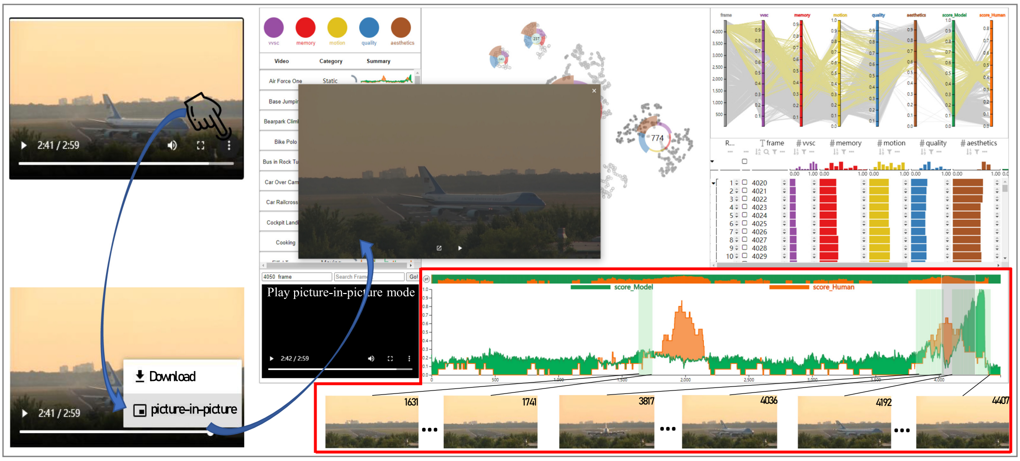


Fig. 4. In the left part, users could browse videos with any size resolution in picture-in-picture mode. In the right red box of the figure, the selected video summary is highlighted in the sequential frames view.

by the model. The sequential frames view (Figure 2(D)) consists of two parts that show the temporal distribution of the frame information of the selected video. The first part is a pixel-bar chart (Figure 2(D_1)) that displays the temporal distribution of the five descriptive features as well as the model results, while the second part is a Playfair-style chart [57] (Figure 2(D_2)) that presents the difference between results obtained by saliency generation and manual annotation results.

6.1 Design Goals

We derive the following design principles to design visualization techniques that help to address the research questions.

G1 **Provide an overview of scores by manual annotation and saliency models in video collections.** In various categories of video, it is necessary to summarize the scores of manual annotation and saliency models, which could help summarize the entire video collections and relax the issues raised in R2. An overview could help users identify videos of interest, and guide users for further effective exploration.

G2 **Multi-level visual exploration and analysis over a single video.** Our work follows basic principles in visualization design (i.e., overview first, zoom and filter, then details on demand), which has been widely used in the study of complex data. First, we aim to provide users with a summary of the video collections so they could identify the video of interest. After selecting a video, users could further explore event-level information to analyze the connections between adjacent frames. Users could drill down to the frame level to perform detailed analysis within an event of interest, which is important for handling R5 and R6.

G3 **Support joint multi-view analysis within video, event, and frame levels.** A visual analysis system should contain multiple views, which is closely coordinated with friendly interactions to perform an in-depth exploration of visual analysis tasks. Therefore, the system should enable users to interact with any view, and changes should be made to all other views to explore the data from different perspectives (R1–R6). Video playback should also be supported, and the presented data in different views should be linked to the raw video content.

G4 **Reduce visual clutter of video frames presentation.** In the visualization of video frames, as the number of frames increases, the overlapping among nodes that represents frames could be severe. This problem could easily reduce task performance or lead to misleading information. Providing certain sampling algorithms to extract keyframes and optimizing dimension reduction results may help reduce visual clutter (R3 and R6).

G5 **Video content viewing under interactive guidance.** Although the video view supports free playback, it is unguided and time-consuming. Therefore, experts recommend that it is necessary to associate with the video view through interactive brushing in the sequential frames view with guidance information. Thus, the video view could provide more indications and purposeful exploration (R1 and R5).

6.2 Video View

The video view is the first part that provides users with initial exploration. During the design process, we found it necessary to provide intuitive guidance for users to explore without introducing a cognitive burden. Therefore, we present this part by following the principles of simplicity and intuition.

As shown in Figure 2(A), the video view is divided into three parts. The top of the part (Figure 2(A_1)) is a legend that shows our color scheme for each feature to help users understand the color combinations between colors and features. The middle part (Figure 2(A_2)) is a video list, which is also divided into three columns, i.e., video title, video category, and video summary. The video title is the name of the video, and the video category is what type the video belongs to, namely, moving, egocentric, and static. The third summary column consists of two parts. The glyph on the left represents the performance (measured by F-measure score), which is indicated by the length of the arc. On the right, a Playfair-style chart is used to represent the difference between the saliency score produced by the model and the manual annotation score. The X-axis represents the length of the video (i.e., frames), and the Y-axis stands for the score value. This gives users a quick overview to help them locate the video segment of interest (R2).

After selecting a video segment of interest, the original video is displayed at the bottom of the video view (Figure 2(A_3)), allowing users to view the detailed video content. In this view, users could drag and drop the timeline of the video to browse the raw video with any size resolution in picture-in-picture mode, as shown in the left part of Figure 4. Together with the sequential frames view, users could further utilize the provided interaction to find corresponding video content.

6.3 Projection View

The projection view is inspired by expert advice (R6) and acts as a major component.

We use the t-distributed **stochastic neighbor embedding (*t*-SNE)** [58] projection algorithm to map all frames of the video. Each frame is described by a feature vector, which is constructed in Equation (5). The vectors are further mapped to two-dimensional space since it is suitable for embedding high-dimensional data into two-dimensional space to obtain a quick overview of large data. The closer two nodes are in two-dimensional space, the more similar they behave in the high-dimensional feature vector space. As a result, each frame is mapped to a two-dimensional dot, and similar frames are placed nearby.

$$V_c = [S_{vvsc}, S_{memory}, S_{motion}, S_{quality}, S_{aesthetics}], \tag{2}$$

where S_{vvsc}, S_{memory}, S_{motion}, $S_{quality}$, and $S_{aesthetics}$ represent five features, i.e., *visual-verbal saliency consistency*, *memory*, *motion*, *quality*, and *aesthetics* of each frame, respectively.

Initially, each frame is represented by a dot, but this does not give users any visual indication and introduces severe visual clutter. Therefore, according to the requirements of the analysis task

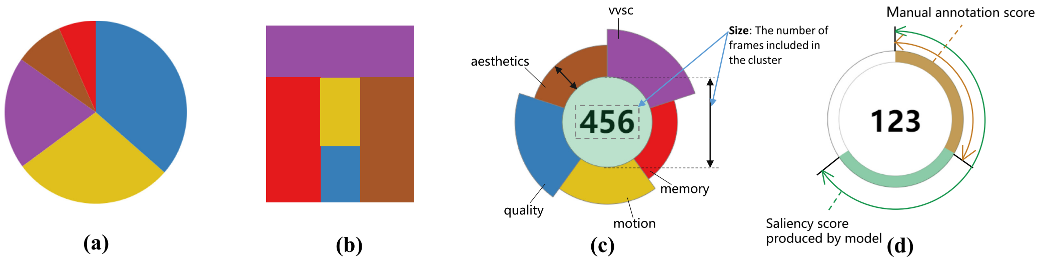

(a)
(b)
vvsc
aesthetics
Size: The number of frames included in the cluster
456
memory
quality
motion
(c)
Manual annotation score
123
Saliency score produced by model
(d)

Fig. 5. Our alternative and final glyph designs. (a) A nested pie chart design. (b) A treemap-based design. (c) An event-based projection with a glyph design showing feature information of the frames in the cluster. (d) Glyph design of a single frame, starting from the top of the ring; orange indicates manual annotation score, while green indicates saliency score produced by the model. The number inside the circle indicates the current frame ID.

(R4–R6) and our design goal (G2, G4), we use a color gradient to characterize the chronological order of the frames. The brighter gray value indicates initial frames, while the darker gray value stands for the later frames. Thus, users could obtain an overview of how temporal patterns are presented in the projection view. We further use the DBSCAN algorithm to cluster the two-dimensional vectors obtained by dimensionality reduction to obtain multiple clusters. We regard each cluster as an event in the video since we find that the frames in a cluster are mostly continuous and there is a certain similarity between adjacent ones. After reviewing corresponding original video clips, this speculation is further verified. In addition, we have added a new glyph design that provides users with an overview of the features of each cluster to facilitate comparison and selection among clusters.

In the design process, we considered several alternatives as shown in Figure 5. Though pie chart design (Figure 5(a)) could present the descriptive feature score completely, it will introduce additional visual clutter when overlaying the number of frames on it. As for the design based on treemap (Figure 5(b)), it may cause visual confusion and lack of legibility.

Our final design (Figure 5(c)), which combines the advantages of circle and sector, consists of two parts. The internal circular radius encodes the number of frames in the cluster. The larger the radius, the more frames the cluster contains. Color is used to encode the difference between the model and the manual annotation in the cluster. The deeper the green is, the higher the saliency score produced by the model is than the manual annotation score. The deeper the orange, the higher the manual annotation score is than the saliency score produced by the model.

The five equally divided circular sectors are used to encode the five features extracted from the frame image. More specifically, the color is used to represent the name of the feature, and the radius is used to indicate the size of the descriptive feature score. The larger the radius, the greater the average descriptive feature score for all frames in the cluster.

Considering expert advice, it is necessary to drill down to frame level in this view. However, since the current view covers all frames of the video, there are too many nodes in the limited screen. The overlapping between nodes is severe, which causes visual confusion, and the dot containing only the time information is not enough to provide more visual information for users. To better utilize our system in many frames, we have optimized the projection view. For each cluster, we propose a sampling algorithm to sample the frames in the cluster to enhance the visibility of important frame information.

We combine three customized indicators to sample the frames in the cluster, which are *importance indicators, influence indicators, and stability indicators.* The calculation of the indicator's formula is as follows.

$$In_{importance} = (S_{hum} + S_{mac})/(S_{hum} + S_{mac})_{max}, \tag{3}$$

$$In_{influence} = (Dis_{bf} + Dis_{af})/Dis_{max}, \tag{4}$$

$$In_{stability} = (1/|S_{hum} - S_{mac}|)/(1/|S_{hum} - S_{mac}|)_{max}. \tag{5}$$

Among them, S_{hum} represents manual annotation score, S_{mac} represents saliency score produced by the model, Dis_{bf} represents the geometric distance between a frame and the previous frame, Dis_{af} represents the geometric distance between a frame and the next frame, and Dis_{max} represents the value of the longest geometric distance between two frames in the cluster.

We use the distance between a frame and the two frames before and after it to characterize the influence. If the distance between a frame and the previous frame or the next frame is relatively large, it means that the similarity of the two frames tends to be low. Such a frame could be a video scene switch or content mutation, which should be retained.

We further employ a higher model score or manual annotation score to characterize the importance. If the saliency score produced by the model or manual annotation score of the frame is relatively high, such a frame is likely to be a keyframe in the video and should be retained. For the *stability indicator*, we use the closeness of the saliency score produced by the model to the manual annotation score to indicate the stability of the frame. If the saliency score produced by the model is closer to the manual annotation score, the frame is more stable. Such frames are often selected as video summary frames and should be retained.

In this article, frame sampling is mainly utilized to reduce severe visual occlusion in the projection view. Since the information among closer frames is basically similar, the sampling part of the frames would not have a prominent negative impact during visual analysis. In the projection view, if the radius of the frame's glyph is x, we take the position of a frame as the center and $2x$ as the radius to keep the frame with the highest importance, influence, and stability in this circle during sampling. Since the selection of different starting frames could lead to different sampling results, we employ a full quadtree to calculate the density of each node. First, starting from the node containing the largest number of frames, and the frame closest to the geometric center of the node as the start frame. Then traverse in sequence according to the distance between frames from small to large. Finally, start from the node with the next highest number of frames, and so on. We have found that when the length and width of the nodes are four to five times the diameter of the frame's glyph, the effect of stopping the process of splitting the grids of quadtree intends to be optimal after experiments, given the length of our projection view is 800 px and the width is 640 px.

The projection view, as well as the sampling process, are shown in Figure 6. It can be seen that the visual clutter and occlusion among frames in the projection view have been alleviated to a different extent. The number of frames is reduced while the relative position among frames is maintained. In addition, the information of keyframes is preserved while reducing the redundancy of similar frames. Users could also click the frame node in the sampling view to examine the three indicators of certain frames.

6.4 Detailed View

As shown in Figure 2(C), the detailed view is interactively linked to the projection view, providing detailed information and exploration.

In the top half, we provide a parallel coordinate view to display statistics for all frames of the selected video. In this view, each polyline corresponds to score information for one frame in the projection view, including descriptive feature scores (i.e., *vvsc*, *memory*, *motion*, *quality*, and *aesthetics* scores), saliency score produced by the model, and manual annotation score, as shown in

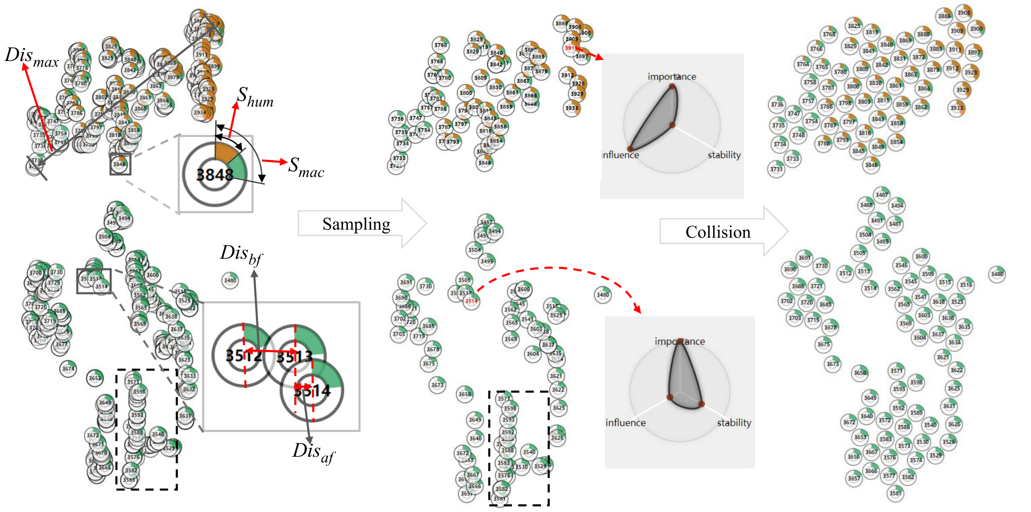


Fig. 6. Sampling process of the glyph nodes in the projection view. Based on three customized indicators, our sampling algorithm could alleviate the severe occlusion issue among the larger number of frames, while preserving keyframe information in the projection view.

Figure 2(C_1). Users could select a specific sequence whose attribute value is within a certain range by brush on each coordinate, and the polyline selected by the brush is highlighted. Other polylines are grayed out and remain as the background. In the lower part, we provide a LineUp [59] view (Figure 2(C_2)) to show the statistics of the cluster and the differences between different clusters. In this view, there are eight columns, showing the number of frames, five descriptive feature scores, saliency score produced by the model, and manual annotation score. In the dotted box at the head of the view, the distribution of scores in each scored segment is provided. At the same time, it also supports typical interactive functions such as filtering and sorting.

6.5 Sequential Frames View

The sequential frames view provides a novel perspective for user exploration in the temporal dimensions.

To reveal the overall temporal trend of the scoring situation, and explore the detailed video content and model effects, together with the advice of the experts (R3–R5), we propose a sequential frame view to provide users with a temporal exploration analysis. In this view, we use two sub-views to display all the information of the video comprehensively by switching the view.

As shown in (Figure 2(D_2)), the Playfair-style chart reveals the score distribution and difference between the saliency model and the manual annotation. The X-axis represents the length of the video, and the Y-axis represents the score per frame. The green polyline indicates the temporal distribution of our saliency score produced by the model, and the orange polyline indicates the temporal distribution of the manual annotation score. The light-orange area indicates that the manual annotation score is greater than the saliency score produced by the model during this period, while the light-green area indicates that the saliency score produced by the model is greater than the manual annotation score during this period. Users could switch to the feature score view by clicking the toggle button, which is presented in (Figure 2(D_1)). This view shows the overall temporal trend of the five features of *vvsc*, *memory*, *motion*, *quality*, and *aesthetics*. The color corresponds to the legends illustrated in (Figure 2(A_1)).

At the top of the two subviews, we also provide an additional pixel bar to represent the proportional change in saliency score produced by the model and manual annotation score over frames. The color of the bar in the pixel bar view represents diverse types of scores, and the height indicates the value of the size, i.e., green indicates the saliency model and orange indicates manual annotation.

6.6 User Interactions

Our system VSumVis supports a variety of interactions, giving users powerful visual analysis capabilities. The four views provided in the system are well-linked together. Here, we summarize the interactions used in our proposed system.

Basic Interaction. When users click on the video of interest in the video view, the corresponding video is selected and other views are updated accordingly. In the projection view, when the users' mouse moves over the cluster glyph of interest, the corresponding frame in the cluster will be selected and highlighted. When the mouse is clicked, the selected cluster will be displayed on the current page in an unobstructed layout. In the detailed view, the details of all the frames in the cluster are displayed. The re-layout node also supports click operations and then displays the frame information of the node in the detailed view.

Brushing. When users brush the sequential frames view, the detailed information of the frame in the selected frame range and the original video content of the corresponding period in the video view are displayed in the detailed view. In addition, the parallel coordinate view in the detailed view also supports the brushing operation.

Searching and Sorting. Searching and sorting in the video view and detailed view, we add search and sort interactions to allow users to quickly find frames of interest. Users could perform specific frame searches based on instructions from other views and sort various scores according to specific criteria.

Lassoing. We have added lasso tools to support free exploration. In the projection view, users could use the lasso tool to select the frame nodes of interest, and in the detailed view, the details of the frame of the selected area are displayed.

7 EVALUATION

This section presents two case studies and a user evaluation to demonstrate the usability and effectiveness of VSumVis.

7.1 Case Studies

We use the SumMe dataset to visually explore the diagnosis mechanism and select the video *Jumps* and the video *Air Force One* to perform a detailed analysis of the single video. This section shows how users could perform diagnosis tasks on saliency models and video content through a visual analysis system.

Visual Diagnosis Based on Video Content. This case study shows how VSumVis could help users analyze video content and quickly locate keyframes. We chose the video *Jumps* as the primary unit of analysis to explore its diagnosis methods.

VSumVis allows users to quickly browse summary information for all videos in the SumMe dataset and accurately select the videos of interest. The video view is shown in Figure 2(A). It shows the names, categories, and summary information for each video in the video collections, as well as raw video playback. The summary information shows the scores of the two scoring schemes along with the specific trend of the video frame. For example, in the summary information of the video *Jumps* (Figure 7(a)), it is apparent that the two-color lines of the time period T (see Arrow T in Figure 7(a)) tend to be relatively high and relatively similar. It indicates that the video *Jumps* both

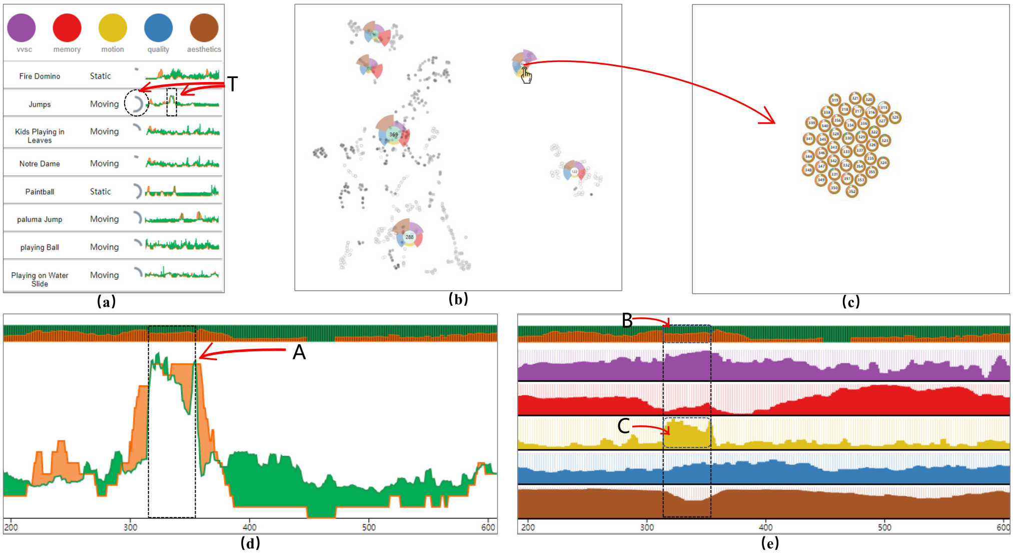


Fig. 7. (a) Select a video of interest based on the summary information in the video collections. (b) Two-dimensional projection of all frames of the *Jumps* video, and corresponding cluster information. (c) Re-layout of the frames in the cluster. (d) The area of the keyframes in the video *Jumps* with Playfair-style chart. (e) Investigation of the keyframes with visualization of the five descriptive visual features based on the pixel bar chart.

have a higher score with respect to the saliency model and manual annotation during this time period, which also implies that the video in time period may contain more keyframes. Effective real-time contextual exploration of video content through video views promotes user confidence in the summarized information.

VSumVis allows users to explore detailed content in the video. We first select the video of interest (e.g., *Jumps*) in Figure 7(a), which is a Moving type video with a duration of 39 seconds (a total of 940 frames). The projection view (Figure 7(b)) shows the two-dimensional distribution of all the frames of this video. Observing the six clusters formed by clustering in this view, we could boldly assume that the content of the video could be roughly divided into six events. The diameter of the circle in the glyph of each cluster indicates the duration of the event (i.e., number of frames). For example, the circle of the glyph in the middle of the projection view is the largest (see Arrow D in Figure 7(b)), and we could find the cluster is composed of 369 frames through examination. By observing the video view, the main event in the video is a group of people celebrating. In addition, observing the sector of the periphery of the cluster glyph, it is also possible to view the average score of the features of the video frame image composed of the current event. We could see that the average score of the cluster's *memory*, *quality*, and *aesthetics* features are relatively high.

In addition, in the projection view, we observe the specific events of the video content (such as the pull of the lens, the man sliding from the slide to the sky, landing in the pool, and the celebration phenomenon). However, such analysis stays at the event level; the exploration of the entire video content is not careful enough. To further analyze the composition details of the event, we click on the current cluster glyph, and the original projection view area is switched to the current cluster's dot plot (Figure 7(c)). It shows the frame composition of the clicked cluster. Each node has corresponding frame information and is bounced off by the collision detection mechanism under analog force, thereby avoiding visual overlapping and information confusion. We could see that each node is tiled in the view area without overlapping each other. By linking to the detailed view, users could see the specific information and distribution of each frame.

VSumVis could position keyframes and perform analysis of video content with sequential frames view. Figure 7(d) shows the temporal distribution of the video *Jumps*. We could see that during the period of 315th–356th frames (see arrow A in Figure 7(e)), both the green and orange polylines have reached a peak. This means that the saliency model and manual annotation give a fairly high score for the image frame in this period. To further confirm our observation, we studied why the saliency model has a high score, and whether there is a certain correlation between the score and the input features of the model. By clicking the button (Figure 7(e)), we switch the current page to the feature view. When examining the temporal trend of five descriptive feature scores during that period, we found that the scores of *vvsc*, *motion*, and *quality* features are generally higher than those of other periods. The *motion* feature is particularly high in this period and behaves extremely low in other periods. The region's *motion* feature score (see arrow C in Figure 7(e)) reaches a fairly high level, and we suspect that this may be the main factor for the high saliency model score. After further observing the content of the raw video, we find that the people in the video are playing air taxiing, which means that the *motion* feature is well-extracted from the video. This confirms the reason for the high score of the model. In addition, we observe that the ratio of green and orange in the pixel bar chart is close to 1:1 (see arrow B in Figure 7(e)), indicating that the scores of manual annotation and the scores of saliency models during this period are remarkably high, and the difference between the two is small. We speculate that this period is a period containing many keyframes.

Visual Diagnosis Based on Saliency Models. The effects of the diagnosis mechanism in VSumVis have been verified through multiple examinations of the machine learning model on SumMe data.

VSumVis could help users with a better understanding of the model performance and model hysteresis. The analysis of the *Jumps* video helps to clarify the hysteresis phenomenon in the model as shown in Figure 8. In the left Figure 8(a), the manual annotation score climbs quickly from the 289th to the 308th frame with a sharp rising curve. The corresponding images of the two frames in the video are shown in the top-right, where the person in both of the images is going to slide out from the wooden chute. However, the saliency score curve in the same time period looks gentle and then rises rapidly in five frames (from the 312th to the 317th frame). The bottom-right plots in Figure 8 show the images of the two frames, i.e., the 312th frame, and the 317th frame, in which the person slides out from the wooden chute. We define such a phenomenon as model hysteresis, which means the reaction of the saliency score curve is always a bit behind the manual annotation score curve. We suspect it is the prior knowledge of mankind that leads to model hysteresis. While annotating the video, one would score it high before a scene moves to its peak moment. The machine learning model lacks a perception of actions. Thus, only if a frame has significant features, may it mark the frame with a high score.

Similarly, we observe a greater hysteresis in the *air force one* video which depicts the landing of the airplane. In Figure 8(b), from the 3,954th frame to the 4,054th frame, the airplane is moving from a straight line and starts to turn, which is a coherent action. From the model's perspective, the five features of the image in this period do not change much. However, humans are able to know in advance that "it is going to turn right away," so they start to pay attention ahead of the computer. From the 4,070th frame to the 4,211th frame, during the turning action of the airplane, the computer-calculated features change and the *motion* score begins to rise. After observing the video collections, we found that in the videos of the static type, the probability of this hysteresis phenomenon is the largest, and ranks second in the moving type video and third in the egocentric type video. The result is in line with our understanding. Regarding the static type video, since

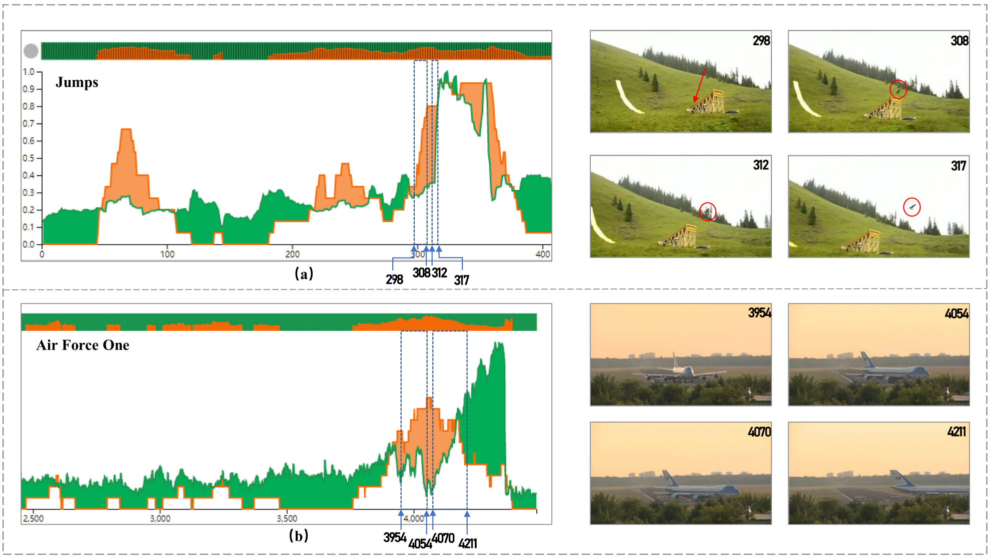


Fig. 8. Hysteresis phenomenon in the model. (a) Regarding the video *Jumps*, the model has a hysteresis phenomenon in the time period when people slide out from the wooden chute. (b) Regarding the video *Air Force One*, the model has experienced a hysteresis in the time of period when the airplane turns.

the camera position is fixed, the video content tends to be stable. Thus, our model could capture the features of each frame. Regarding the egocentric type videos, the camera is less stable when shooting the egocentric type of video as the shooting device is fixed on the photographer's head, so the picture is unstable during feature extraction, and subtle changes are not easy to capture in adjacent frames. Our visual analysis system provides a new perspective to study the performance of the saliency model by exploring the possible hysteresis between model output and manual annotation.

VSumVis allows users to quickly understand the relationship between features and model output. When we observe the parallel coordinate view, we find that there is significant differentiation between the saliency score produced by the model and the motion score. As shown in Figure 9(a), we speculate whether there exists a connection or not. Thus, we perform navigation in the sequential frames view, and when sliding to the 300th to the 323rd frame (see arrow A in Figure 9(b)), the *motion* score becomes higher. To confirm our speculation, we click the button to switch to the feature view; the trend of the *motion* is relatively high (see arrow B in Figure 9(c)). We shift our eye to the detailed view, and we find that the scores for *motion* and saliency model begin to diverge, and the score becomes higher. We speculate that the increase in the saliency score produced by the model at this time is mainly due to the increase in the *motion* score. We then view the corresponding image frame (Figure 9(d)), and it verifies the changing process (a person slides from the wooden chute to the sky).

VSumVis could provide an explanation for the large difference between the saliency score produced by the model and the manual annotation score. In the period from the 1,830th to the 2,130th frame (see arrow A in Figure 10(d)), it could be seen that the orange polyline is much higher than the green polyline, and in the pixel bar, the orange portion accounts for a considerable proportion. An explanation during this time period could be the following: human observations believe that the series of frames is important, and the manual annotation score is high. However, the model predicts the score to be kind of low. By examining the raw video, it is the scene when the airplane

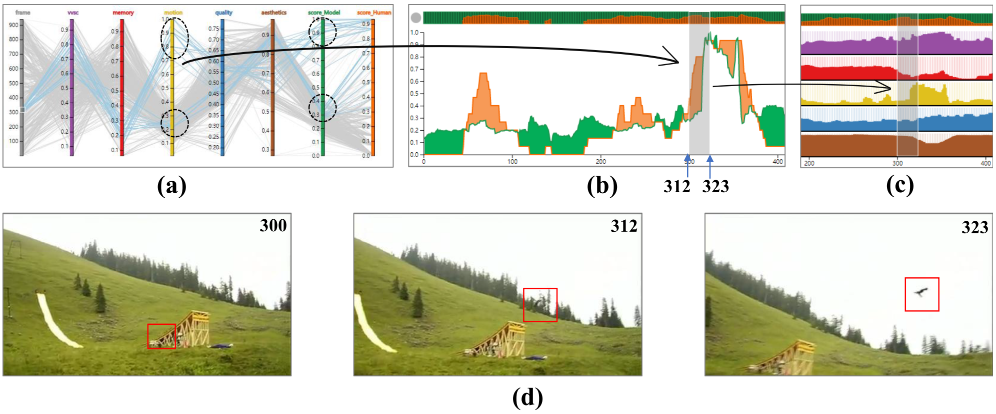


Fig. 9. Exploring the potential relationship between the saliency score produced by the model and corresponding features. (a)–(c) Exploration process from parallel coordinate view to sequential frames view. (d) Verification from the raw video frames.

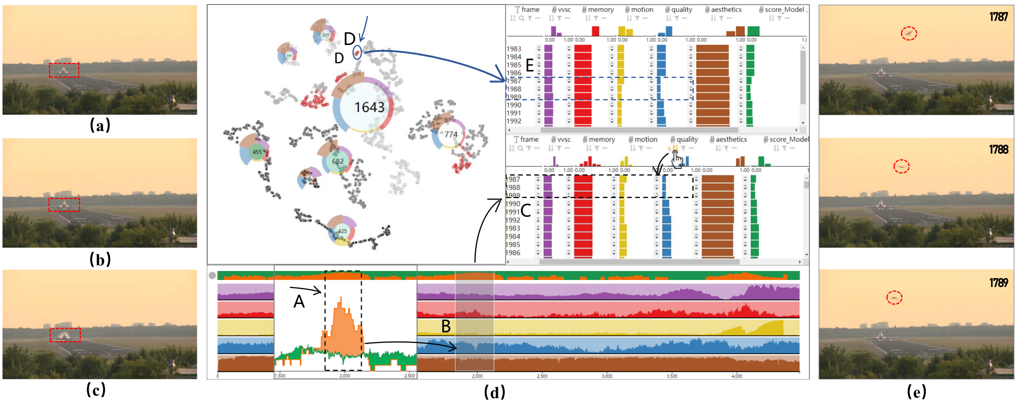


Fig. 10. (a) The video shows that the airplane is about to land. (b) The airplane is landing. (c) The plane is already landed. (d) Visual analysis process for anomaly frame fast located. (e) The anomaly raw video frames.

is landing (Figure 10(a)–(c) are the image frames of the upcoming landing, landing, and landed, respectively). After landing, the dust generated by the landing of the airplane caused visual attention. The information was attracted by human vision, and the model did not capture the corresponding information in time. We speculate that this is because there is no corresponding feature to extract the information of these changes, resulting in a large deviation between model prediction and manual annotation. Combined with the hysteresis, the model is not consistent with human observation when the airplane is about to land.

VSumVis could quickly locate the location of the anomaly frame and explore the cause while offering a certain explanation. Similarly, in the period from the 1,830th to the 2,130th frame, we find that the saliency score produced by the model of the individual frames tends to be an anomaly. We speculate that this is due to a sudden change in a feature, thus we switch to the feature view (see Aarrow B in Figure 10(d)). It is found that during this period, the *quality* score suddenly decreases, indicating that our speculation is in the right direction. To find out the anomaly frame for an in-depth analysis, we select the corresponding period in the feature view. In the detailed view, it is verified that the *quality* feature has a low score. We sort the columns of *quality* from small to large

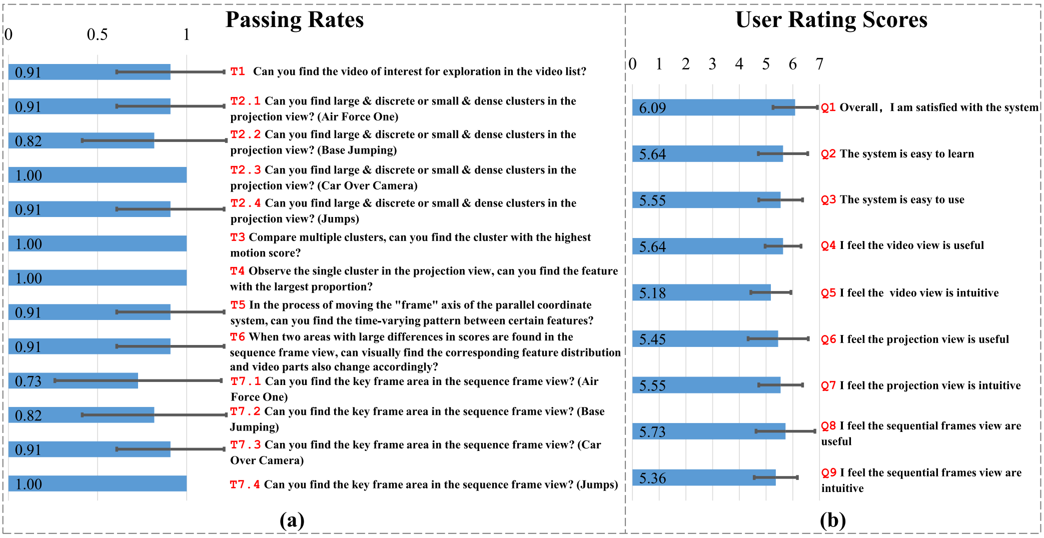


Fig. 11. (a) Evaluation tasks and analysis of the passing rates. (b) Analysis of the ratings: all questions received an average rating greater than 5, which was very encouraging.

in the LineUp view and find the position of the anomaly frames, which are the 1,987th, 1,988th, and 1,989th frames (see arrow C in Figure 10(d)), respectively. On the other hand, we want to verify the validity of our t-SNE algorithm. According to the algorithm's dimensionality reduction, the anomaly frame should deviate from other frames in the region. By using the lasso to select the area that deviates from the other frames in the projection view (see arrow D in Figure 10(d)), we find that the above anomaly frame is in the selected area of the lasso in the LineUp view (see arrow E in Figure 10(d)). This confirms our speculation of the anomaly frames. To further analyze the cause of the anomaly frame, we look up the video segment where the frame is located in the video view (as shown in Figure 10(e), three anomaly image frames are excerpted), and find that it may be dust or birds caused by airplane landing, which reduces image quality to a certain extent.

7.2 User Study

Next, we conducted a laboratory study to examine how well users can use our VSumVis to explore and gain insights into video content as well as the summarization information. The study also aimed to find any usability issues for improving the system.

7.2.1 Participants. We invited 11 students (6 male, 5 female) between the ages of 19 and 27, to evaluate the proposed system. They were not the co-authors of the article. They comprised both undergraduate and postgraduate students, with backgrounds in computer science, automatic, electrical engineering, and machine learning. All of them had certain knowledge in the visualization or machine learning field. Thus, they are competent for analysis tasks. The participants were identified as P1–P11, respectively. The use study was conducted based on the SumMe dataset.

7.2.2 Tasks and Procedure. At the beginning of each study, we demonstrated every detail of VSumVis and provided a tutorial for each participant. Once each participant is familiar with the system, the dataset is loaded, and the participant is asked to complete a series of tasks (Figure 11(a)) to explore the video content using VSumVis. These tasks appropriately cover various visualization components. If necessary, participants can request further explanations of the functions of VSumVis. When the participants completed the task, they were asked to rate various aspects of VSumVis on a Likert scale (from strongly disagree to strongly agree which is represented with 1–7). Among them, according to the specified four videos, T2 and T7 are divided into four subtasks. Those tasks (shown as T1-T7 in Figure 11(a)) were designed to utilize all visualization views and reflect the complete set of proposed requirements in Section 3.2. Specifically, T1 correspond to R1; it is used to test how users find interesting videos in the video collections. T2, T3, and T4 correspond to R6; many similar frames may form clusters, and we design glyphs to distinguish them. T3 and T4 test the effectiveness of glyphs. As for T2, the frames in the dense clusters are usually similar, which is used to examine whether users could observe similar patterns in the projection view. T5 and T6 correspond to R1, R4, and R5; these tasks would test how users discover patterns that existed in the model and how they explore the dynamic association between frames, input features, manual annotation, and model output from a system view. T7 corresponds to R3; keyframe areas often are areas where significant anomalies occur; it could help users find them. The task-based assessment could help participants better understand the system, thereby ensuring that they can obtain more objective and comprehensive feedback on the individual functions of the system and the entire system from the participants. At the end of each study, users are free to explore the system. We conducted post-study interviews to collect feedback from participants. The user study for each participant lasted approximately 40 minutes and was conducted on a 23.8-inch monitor.

The above evaluation received an average pass rate of 90.9% with a standard deviation of 0.24. The pass rates of T3 and T4 are both 100%, which proves the effectiveness of VSumVis in discovering interesting videos and cluster glyph design. The pass rate of T7 is one of the lowest. During the interview, we learned that most of the participants had doubts about the definition of keyframes. Since the model score is a fitting result of the manual annotation score, there may exist a divergence between the saliency score produced by the model and the manual annotation score. Therefore, in the sequence frame view, when certain manual annotation scores are relatively higher, while the saliency scores produced by the model are lower, participants may hesitate when deciding. It is because of these differences that users can pay attention to which frames model may have problems and then analyze which aspects of the model itself can be improved for these frames.

7.2.3 Questionnaire. Participants completed a questionnaire with nine questions. Figure 11(b) shows the questions and average user ratings. For each question, users are asked to rate the satisfaction of the system. In the overall satisfaction score of the system, users have the highest satisfaction with the system (reaching the highest score of 6.09). Based on the results of Q2 and Q3, VSumVis seems easy to learn and use. This is an encouraging result because unpracticed users can complete fairly complex exploration tasks with minimal assistance during the experiment. When considering each component separately, participants believed that the visual performance and interaction of the projection view were intuitive during the exploration process (Q7). However, even though the video view was deemed useful (Q4), participants found the list-based video collection difficult to understand at first glance (Q5). After discussions with the participants, we found that this result was due to the lack of metaphors in their brains. Through demonstrating the use of the video view, they got used to it. We also found that the sequence frame view received the highest score (Q6–7) in terms of usefulness and intuitiveness. After interviewing the participants, we found that they appreciate the real-time interaction with the video content in the process of exploring the difference between manual annotation and model scores. In general, the average score for all questions exceeds 5 points, which is very encouraging.

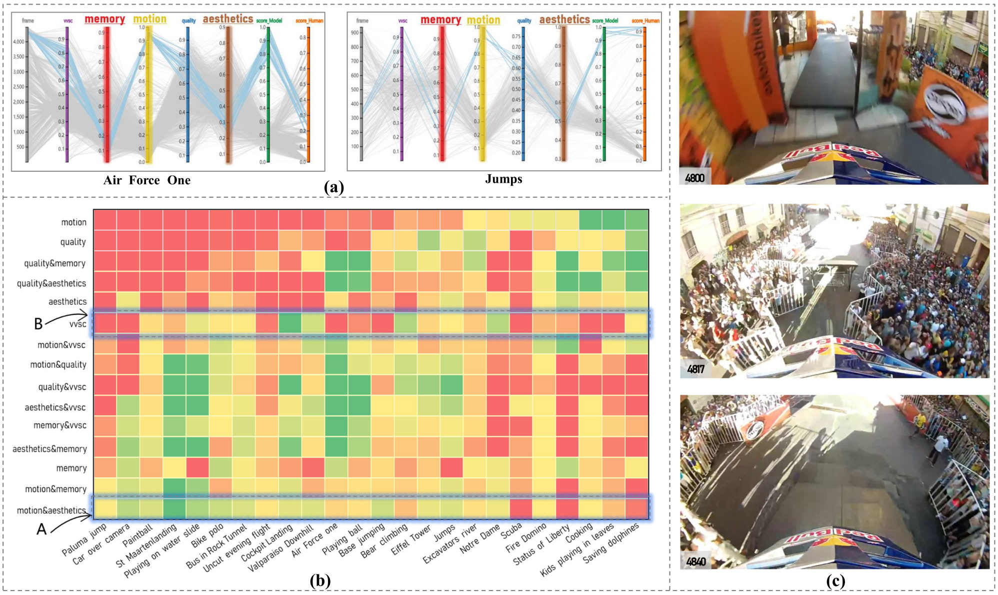


Fig. 12. (a) An abnormal pattern among *motion*, *memory*, and *aesthetics* features is observed in parallel coordinates. (b) *Performance grid view* to demonstrate the overall distribution of the impact by the model terms as well as their interaction. (c) In the video *Valparaiso Downhill*, a lot of scene changes are observed in less than 2 seconds (from the 4,800th to 4,840th frame).

7.3 Experts Feedback

To evaluate the effectiveness of VSumVis, we interviewed one assistant professor (PA) who engaged in computer vision area research for more than 9 years and four postgraduate students. PA also plays a significant role in proposing the saliency generation model for this work, and he is eager to know how the interaction among different features may affect the model performance. Thus, we first describe the visual encoding and user interactions in VSumVis and then demonstrate the patterns observed in the case studies. PA is convinced by the effectiveness of the system and conducts a diagnostic evaluation of the VSumVis himself to find out whether the interaction of the above five descriptive features may affect the model performance or not.

During the exploration process, PA observes an abnormal pattern between several features using parallel coordinates. For example, regarding the *Air Force One* and *Jumps* videos (as shown in Figure 12(a)), the value of *motion* score tends to be relatively higher, while the *memory* and *aesthetics* scores tend to be lower. This pattern could also be found in a bunch of videos of different types. However, the impact of such a pattern varies among different video types, which may lead to unreliable performance if we employ a uniform model (i.e., with the existence of all item interactions) to generate saliency scores for all videos. Therefore, PA speculates whether the term interaction in the saliency generation model with respect to the above features may affect the model performance. Thus, PA considers removing the specific interaction terms in the model through the model manipulation interaction in the VSumVis system to examine the later model effect. The performance of the model output is evaluated with F-measure and visualized with a *performance grid view* to examine the overall distribution of term impact (as shown in Figure 12(b)). Each row represents the model performance (measured by F-measure score) with the inexistence of a specific interaction term, and each column represents a specific video in the SumMe dataset. The color in each grid indicates the model performance. The greener the grid is, the better the performance tends to be compared to the default model, and vice versa. The *performance grid view* supports row and column reordering under different criteria to better reflect the overall patterns in the video collection. For example, in Figure 12(b), rows with the higher number of redder grids are placed upwards, and vice versa.

Through observation of the *performance grid view*, PA finds that removing the interaction terms of *motion* and *aesthetics* results in better performance of the salience generation model (see arrow A in Figure 12(b)), which arouses his interest. Thus, PA explores the raw frames of the videos in the row of *motion* and *aesthetics*, and finds that most of the videos that exhibit better performance exist in scenarios of significant action changes or lens switching (e.g., Figure 12(c)). He speculates that these scenarios may reduce the degree of *aesthetics*, and further bring down the explainability of the interaction between *motion* and *aesthetics*. Thus, he concludes that for videos with significant action changes, it is suggested to remove the interaction term of the above two terms.

Through further examination of the *performance grid view*, PA locates several other interesting and unusual patterns. Regarding the videos used in the SumMe dataset, the removal of the semantic feature of *VVSC* degrades the model performance for certain videos (see arrow B in Figure 12(b)). However, PA originally believes that adding the semantic feature of *VVSC* should improve the model performance for most videos. Therefore, PA examines the raw frames of the videos in our VSumVis system and finds that there exists many scenes switching and screen shaking in these videos. Thus, he speculates that this situation may lead to the inconsistency between the verbal and visual features in the *VVSC*, which further affects the model performance. To address this situation, he concludes that, for real-world applications, the video summarization methods should first employ certain traditional computer vision techniques such as frame difference or optical flow to calculate the significance of scenes switching/screen shaking, and further determine the inclusion or not of the *VVSC* feature.

During the interview with the experts, we also collect their feedback, which is summarized as follows.

Method Design. This method has received positive feedback in expert interviews. They agreed that this method provides a workflow for video collection analysis and is a good visual analysis of video content. The proposed system uses computer vision technology to capture video information, extract visual and semantic features, and apply saliency models to fit manual annotation. Finally, the complete process is presented through visual analysis. In the visual analysis system, they confirmed that the multi-perspective and multi-level coordinated analysis could inspect the model effect and raw video content. Regarding the methodological recommendations, one expert added that there are currently fewer features extracted, and to capture more video information and content, it is possible to add as many features as possible to perform trial and error tests. Regarding the machine learning model of the application, experts suggest that when the number of features increases, it is desirable to use a more complex deep learning model due to the sharp increase and complexity of the data volume, which could make the model more accurate.

Visual Design. All users appreciate our visual analysis system. They believe that the multi-level (video, event, and frames) visual analysis method has greatly helped them reduce cognitive pressure in the process of exploring video collections. The professor likes the design of the cluster very much, he added. “This design is novel and valuable to me. It could dig down into frame-level information.” The layout based on collision detection algorithms has also been appreciated by him. Regarding the improvement of visual design, an expert suggested that in the sequential frames view, showing the ratio of saliency score produced by the model to manual annotation scores may be more conducive to temporal analysis.

Usability. All the users are convinced by the practicality and effectiveness of the system, and they all expressed the hope that the system will become public. The professor mentioned that the model exploration process from overview to detail is reasonable, and user interaction is smooth and easy to master. All students appreciate multi-view coordination and multi-level analysis to study the hidden insights behind the observed differences. With respect to usability improvements, three students suggested that the system should support the integration of multiple data types, such as visual representations of other multimodal data like text, images, and audio.

8 DISCUSSION

Our work addresses important challenges in the study of video summarization. While our design brings together multi-feature pattern analysis and multi-level exploration into an interactive visual analytics system. The VSumVis system supports a deep analysis of saliency models and manual annotation video summaries, which is a fundamental work for studying inter-relations of video summarization and machine learning models.

The effectiveness and flexibility of our system are demonstrated in both machine model diagnosis and user guidance. It helps researchers in understanding the applicability of machine learning models and the effect of each descriptive feature in video summarization. While users are investigating the manifestations underlying descriptive features, they may find some useful insights for creating new features that are better able to describe a video.

The visual analysis progress that allows interactive exploration of the intrinsic relationship among saliency models, features, and raw video content would be easily extended to other research topics, especially the research with respect to multimedia visualization. For example, one could extend our work to audio analysis and further gain additional benefit with partial modification on system components.

Domain experts are fully involved in the visual design process, starting from requirement identification to system evaluation. With the ordinary understanding that the visualization system should be concise and easy to use, multiple coordinated views are proposed in the VSumVis system which makes full use of the screen and allows users to efficiently explore the videos. Though the system works with fair manifestation and domain experts all agreed on the usefulness of the visual analysis system, they provided us with suggestions about further improvement of the tool. One suggestion would be to add a widget that integrates more advanced features that are better able to allow users to intuitively see and follow the changes. Another one would be to add new components that allow users to interactively tune parameters.

Scalability. In general, there are two scalability issues. First, there are five features used in our model; if there are more features, it may limit space of the screen. Too many features often lead to inevitable display incompleteness. Because we use pixel-based feature visualization, this form of visualization is essential to alleviate the scalability problem as much as possible, while preserving useful information. In addition, we plan to explore more interactions, such as paging, sliding, and so on. Secondly, if the number of single video frames is too large, it may cause visual confusion in the projection view. Thus, we reduce information redundancy by sampling keyframes, and then combine with collision detection to achieve an occlusion-free result, hence users could focus on discovering valuable frames. If the number of these frames is still large, users could also adjust them through sampling parameters.

Generalizability. Although the goal of our work is to understand the model and analyze the video collections, it is not limited to the model in our article. It is well-known that a feature-based model is commonly seen in the machine learning field, we speculate that our visual analysis method and visualization components could be adapted to other feature-based models. In our design process, the extracted feature-based visualization tasks could cover a wider range than ordinary systems and guide our design. The three-level analysis (namely, video, event, frame) framework we proposed could be used for similar analysis in other hierarchical model analysis. One example is to explore the characteristics and flow of traffic situations. We could extract each road condition and event (e.g., congestion, unblocked, emergency) from the urban road network, and extract key features from traffic sections (e.g., human movement, vehicle flow, number of congestion segment, number of traffic lights). These data could be used and represented with some modification on visualization components in our system. In addition, the projection distribution of multi-dimensional data is very commonly used in visual design, but visual clutter is a common challenge in visualization. The rule-based sampling algorithm we proposed provides a feasible solution for relaxing the issue of overlapping nodes. Most of the current sampling algorithms do not consider realistic constraints for the problems that need to be solved. The sampling algorithm we proposed is based on multiple conditions, which could be more adaptable to solve problems in specific situations. At the same time, these rules could be replaced in other fields, but the method is universal.

Limitation. The limitations of our system remain as follows. (1) Our work relies on the features extracted in the model analysis step. User intervention in feature extraction will inevitably bring uncertainty which further leads to analysis biases. Moreover, there always exists a gap between one's perception and feature representation. A subtle difference in the feature extraction by different users might yield huge analysis heterogeneity. In addition, the fixed descriptive features would never be sufficient to represent a video, especially when meeting different research tasks. A comprehensive feature extraction method based on task requirements would be desirable. (2) Occlusion still exists in the projection view. A 3-minute video generally contains more than 1,000 frames. Since the projection view covers all frames of the video, there will be too many glyphs in a limited screen, which inevitably leads to visual clutter. Even though our sampling algorithm can relax the occlusion problem, there may exist a loss of certain important samples. Designing a more sophisticated algorithm and supplemented by accurate interaction techniques may provide a possible solution. In addition, user interaction could help reduce overlaps in the projection view; however, this issue has not been completely solved. It limits the scalability of our system, especially working on large videos. An advanced visualization technique may be worth further investigation to help overcome this issue.

9 CONCLUSION

In this study, we introduced VSumVis, an interactive visual analysis system that helps users explore the intrinsic relationship among video summarization models, video features, and raw content. Our system consists of multiple coordinated views (video view, projection view, detailed view, sequential frames view) and allows multi-level exploration (video, event, and frame level). The potential for a more effective and efficient investigation of the video summarization model is increased through a synergy between multi-feature examination and multi-level visualization. It is exploited throughout a well-designed process of visual analysis as well as collaboration with domain experts from the computer vision community. Case studies and the presented user study based on the SumMe dataset and interviews with domain experts demonstrate that our system could enable efficient and insightful analysis. Our system could benefit those in the visualization research community by providing a more explainable perspective on the machine learning model and video analysis.

REFERENCES

[1] Muhammad Ajmal, Muhammad Husnain Ashraf, Muhammad Shakir, Yasir Abbas, and Faiz Ali Shah. 2012. Video summarization: Techniques and classification. In *Proceedings of the International Conference on Computer Vision and Graphics*. Springer, Berlin, 1–13.

[2] Sachan Priyamvada Rajendra and N. Keshaveni. 2014. A survey of automatic video summarization techniques. *International Journal of Electronics, Electrical and Computational System* 2, 1 (2014).

[3] Philippe Aigrain, HongJiang Zhang, and Dragutin Petkovic. 1996. Content-based representation and retrieval of visual media: A state-of-the-art review. *Multimedia Tools and Applications* 3, 3 (1996), 179–202.

[4] Wei Ren and Yuesheng Zhu. 2008. A video summarization approach based on machine learning. In *Proceedings of Intelligent Information Hiding and Multimedia Signal Processing*. IEEE, 450–453.

[5] Ying Li, Tong Zhang, and Daniel Tretter. 2001. *An Overview of Video Abstraction Techniques*. Technical Report. HP Laboratory.

[6] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Proceedings of the European Conference on Computer Vision*. Springer, 505–520.

[7] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1059–1067.

[8] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 7582–7589.

[9] Yongjin Liu, Minjing Yu, Qiufang Fu, Wenfeng Chen, Ye Liu, and Lexing Xie. 2016. Cognitive mechanism related to line drawings and its applications in intelligent process of visual media: A survey. *Frontiers of Computer Science* 10, 2 (2016), 216–232.

[10] Carsten Rother, Sanjiv Kumar, Vladimir Kolmogorov, and Andrew Blake. 2005. Digital tapestry [automatic image synthesis]. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1. IEEE, 589–596.

[11] Michael A. Smith and Takeo Kanade. 1998. Video skimming and characterization through the combination of image and language understanding techniques. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 61–70.

[12] Yong-Jin Liu, Xi Luo, Yu-Ming Xuan, Wen-Feng Chen, and Xiao-Lan Fu. 2011. Image retargeting quality assessment. *Computer Graphics Forum* 30, 2 (2011), 583–592.

[13] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (2000), 1349–1380.

[14] Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. 2013. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology* 28, 5 (2013), 852–867.

[15] Rita Borgo, Min Chen, Ben Daubney, Edward Grundy, Gunther Heidemann, Benjamin Höferlin, Markus Höferlin, Heike Leitte, Daniel Weiskopf, and Xianghua Xie. 2012. State of the art report on video-based graphics and video visualization. *Computer Graphics Forum* 31, 8 (2012), 2450–2477.

[16] Guodao Sun, Yin Zhao, Dizhou Cao, Jianyuan Li, Ronghua Liang, and Yipeng Liu. 2019. AtoMixer: Atom-based interactive visual exploration of traffic surveillance data. *Journal of Computer Languages* 53 (2019), 53–62.

[17] Aoyu Wu and Huamin Qu. 2018. Multimodal analysis of video collections: Visual exploration of presentation techniques in TED talks. *IEEE Transactions on Visualization and Computer Graphics* 26, 7 (2018), 2429–2442.

[18] Haipeng Zeng, Xingbo Wang, Aoyu Wu, Yong Wang, Quan Li, Alex Endert, and Huamin Qu. 2019. EmoCo: Visual analysis of emotion coherence in presentation videos. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 927–937.

[19] Kuno Kurzhals, Markus John, Florian Heimerl, Paul Kuznecov, and Daniel Weiskopf. 2016. Visual movie analytics. *IEEE Transactions on Multimedia* 18, 11 (2016), 2149–2160.

[20] Hongsen Liao, Li Chen, Yibo Song, and Hao Ming. 2016. Visualization-based active learning for video annotation. *IEEE Transactions on Multimedia* 18, 11 (2016), 2196–2205.

[21] Yong-Jin Liu, Cuixia Ma, Guozhen Zhao, Xiaolan Fu, Hongan Wang, Guozhong Dai, and Lexing Xie. 2016. An interactive spiraltape video summarization. *IEEE Transactions on Multimedia* 18, 7 (2016), 1269–1282.

[22] Tao Mei, Bo Yang, Shi-Qiang Yang, and Xian-Sheng Hua. 2009. Video collage: Presenting a video sequence using a single image. *The Visual Computer* 25, 1 (2009), 39–51.

[23] 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 Transactions on Multimedia* 14, 4 (2012), 1153–1165.

[24] Minjing Yu, Yong-Jin Liu, Su-Jing Wang, Qiufang Fu, and Xiaolan Fu. 2016. A PMJ-inspired cognitive framework for natural scene categorization in line drawings. *Neurocomputing* 173 (2016), 2041–2048. Issue Part 3.

[25] Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. 1999. Video *Manga*: Generating semantically meaningful video summaries. In *Proceedings of the 7th ACM International Conference on Multimedia*. ACM, 383–392.

[26] Amir H. Meghdadi and Pourang Irani. 2013. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2119–2128.

[27] Markus Hoeferlin, Benjamin Hoeferlin, Gunther Heidemann, and Daniel Weiskopf. 2013. Interactive schematic summaries for faceted exploration of surveillance video. *IEEE Transactions on Multimedia* 15, 4 (2013), 908–920.

[28] Benjamin Renoust, Duy-Dinh Le, and Shin'Ichi Satoh. 2016. Visual analytics of political networks from face-tracking of news video. *IEEE Transactions on Multimedia* 18, 11 (2016), 2184–2195.

[29] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5686–5697.

[30] Enrico Bertini and Denis Lalanne. 2009. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*. ACM, 12–20.

[31] Junhua Lu, Wei Chen, Yuxin Ma, Junming Ke, Zongzhuang Li, Fan Zhang, and Ross Maciejewski. 2017. Recent progress and trends in predictive visual analytics. *Frontiers of Computer Science* 11, 2 (2017), 192–207.

[32] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. 2017. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics* 1, 1 (2017), 48–56.

[33] Yafeng Lu, Rolando Garcia, Brett Hansen, Michael Gleicher, and Ross Maciejewski. 2017. The state-of-the-art in predictive visual analytics. *Computer Graphics Forum* 36, 3 (2017), 539–562.

[34] Xun Zhao, Yanhong Wu, Dik Lun Lee, and Weiwei Cui. 2019. iForest: Interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 407–416.

[35] Subhajit Das, Dylan Cashman, Remco Chang, and Alex Endert. 2019. BEAMES: Interactive multi-model steering, selection, and inspection for regression tasks. *IEEE Computer Graphics and Applications* 39, 5 (2019), 20–32.

[36] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2019. RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 299–309.

[37] Mengchen Liu, Jiaxin Shi, Kelei Cao, Jun Zhu, and Shixia Liu. 2018. Analyzing the training processes of deep generative models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 77–87.

[38] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Polo Chau. 2017. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 88–97.

[39] Minsuk Kahng, Nikhil Thorat, Duen Horng Polo Chau, Fernanda B. Viégas, and Martin Wattenberg. 2019. GAN Lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 310–320.

[40] Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. 2017. Understanding hidden memories of recurrent neural networks. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*. IEEE, 13–24.

[41] Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. 2019. SANVis: Visual analytics for understanding self-attention networks. In *Proceedings of International Conference on Visualization*. IEEE, 146–150.

[42] Shouxing Xiang, Xi Ye, Jiazhi Xia, Jing Wu, Yang Chen, and Shixia Liu. 2019. Interactive correction of mislabeled training data. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology*. 57–68.

[43] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 547–554.

[44] Shixia Liu, Changjian Chen, Yafeng Lu, Fangxin Ouyang, and Bin Wang. 2019. An interactive method to improve crowdsourced annotations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 235–245.

[45] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and understanding classifiers with rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 342–352.

[46] Jian Zhao, Chidansh Bhatt, Matthew Cooper, and David A. Shamma. 2018. Flexible learning with semantic visual exploration and sequence-based recommendation of MOOC videos. In *Proceedings of Conference on Human Factors in Computing Systems*. ACM, 1–13.

[47] Sara Di Bartolomeo, Yixuan Zhang, Fangfang Sheng, and Cody Dunne. 2020. Sequence braiding: Visual overviews of temporal event sequences and attributes. In *Proceedings of the International Conference on Visualization*. IEEE.

[48] Qianwen Wang, William Alexander, Jack Pegg, Huamin Qu, and Min Chen. 2021. HypoML: Visual analysis for hypothesis-based evaluation of machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1417–1426.

[49] Joseph F. DeRose, Jiayao Wang, and Matthew Berger. 2021. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics* 27 (2021), 1160–1170.

[50] Haoran Liang, Ronghua Liang, and Guodao Sun. 2016. Looking into saliency model via space-time visualization. *IEEE Transactions on Multimedia* 18, 11 (2016), 2271–2281.

[51] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Transactions on Image Processing* 27, 10 (2018), 5142–5154.

[52] Mengjuan Fei, Wei Jiang, and Weijie Mao. 2018. Creating memorable video summaries that satisfy the user's intention for taking the videos. *Neurocomputing* 275 (2018), 1911–1920. Issue C.

[53] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. 2015. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2390–2398.

[54] Phillip Isola, Jianxiong Xiao, Devi Parikh, Antonio Torralba, and Aude Oliva. 2014. What makes a photograph memorable?*IEEE Transactions on Pattern Analysis & Machine Intelligence* 36, 7 (2014), 1469–1482.

[55] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[56] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1072–1080.

[57] William Playfair. 2005. *Playfair's Commercial and Political Atlas and Statistical Breviary*. Cambridge University Press.

[58] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 2579–2605.

[59] Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. 2013. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2277–2286.

Received October 2020; revised March 2021; accepted March 2021