# Video Visualization and Visual Analytics: A Task-Based and Application-Driven Investigation

Wang Xia, Guodao Sun, Tong Li, Baofeng Chang, Jingwei Tang, Gefei Zhang, and Ronghua Liang

***Abstract*— Video data refers to digital information in the form of a series of frames or images representing continuous motion captured by a video recording device. In various domains such as security, sports, education, and entertainment, a significant amount of video data is generated and stored daily. However, analyzing these videos manually is challenging due to their intrinsic characteristics, including large-scale, redundancy, contextual dependencies, and multimodality. Consequently, researchers have extensively explored visualization techniques to address these complexities. In this investigation, we review the state-of-the-art techniques in video visualization and visual analysis. Initially, we provide an overview of the design space for video visualization and visual analysis techniques. Subsequently, we organize and classify these techniques based on visual analysis tasks and application scenarios, providing detailed descriptions within each category. Drawing upon a comprehensive review of existing research, we provide a critical evaluation and propose potential opportunities for future research. Additionally, we have developed a web-based survey browser for convenient exploration of our created classification framework and the associated scholarly articles (https://zjutvis.github.io/VOVideo/).**

***Index Terms*— Video visualization, video analysis, visual analytics, video data, survey.**

## I. Introduction

THE exponential expansion of video data is a direct consequence of the rapid advancements in digital technology. Various domains, encompassing security [1], [2], [3], sports [4], [5], [6], education [7], [8], [9], and entertainment [10], [11], bear witness to the generation and storage of substantial volumes of video data daily. Consequently, there is an urgent demand for browsing, exploring, and analyzing video data in real world.

Manuscript received 26 June 2023; revised 18 December 2023 and 26 March 2024; accepted 28 June 2024. Date of publication 4 July 2024; date of current version 27 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62422607, Grant 62036009, and Grant 62372411; in part by Zhejiang Provincial Natural Science Foundation of China under Grant LR23F020003 and Grant LTGG23F020005; and in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant RF-B2023006. This article was recommended by Associate Editor C. Yang. *(Corresponding author: Guodao Sun.)*

The authors are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: xiawang@zjut.edu.cn; guodao@zjut.edu.cn; litong@zjut.edu.cn; baofeng.chang@foxmail.com; jwtang@zjut.edu.cn; gopherzhang@163.com; rhliang@zjut.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3423402.

Digital Object Identifier 10.1109/TCSVT.2024.3423402

However, the exploration and analysis of video data encounter two primary challenges. Firstly, the manual review of videos is time-consuming and labor-intensive. Secondly, video data encompasses a wealth of information, encompassing multimodal information and contextual association, thereby presenting challenges for video analysis and exploration [12]. To address the above issues, the field of computer vision has extensively researched and proposed a plethora of methods [13], [14]. Particularly in recent years, automated approaches such as object detection [15], [16], object tracking [17], [18], [19], and image segmentation [20], [21], [22] for video data have significantly enhanced the efficiency and precision of video analysis. Driven by the automation techniques, video analysis is gradually transitioning from manual review to machine-based scrutiny [23], [24], [25].

However, directly applying advanced automated methods to real-world video browsing and analysis encounters limitations posed by following critical issues: (1) the "black box" nature of automated algorithms hinders interpretability and transparency in decision-making processes. (2) machine-generated outcomes are typically presented in a discrete and fragmented manner, lacking meaningful insights that end-users expect for. In conclusion, the cognitive bridge between machines and humans remains incomplete.

To meet the demands of human-centric video browsing and analysis, visualization technique [4], [26], [27], [28], [29], [30] assumes the role of a bridge between automated video processing and human perception. The application of visualization and analysis techniques in video prioritizes human perception and experience, utilizing machines as supportive tools to assist users in viewing video data. The primary objective is to alleviate the analytical burden on video viewers and enhance their understanding of video information. The entire process (as depicted in Fig. 1) starts with the raw video data processing, followed by semantic-level modeling of the video data which abstracts low-level features into high-level semantic information. Finally, through the intuitive visualization of video information and a series of interactive operations, video visualization and visual analysis can be achieved.

There have been some surveys summarizing the research on video visualization and analysis techniques in the visualization community. Earlier surveys examined the literature on video visualization and visual analysis techniques [31], [32], [33].

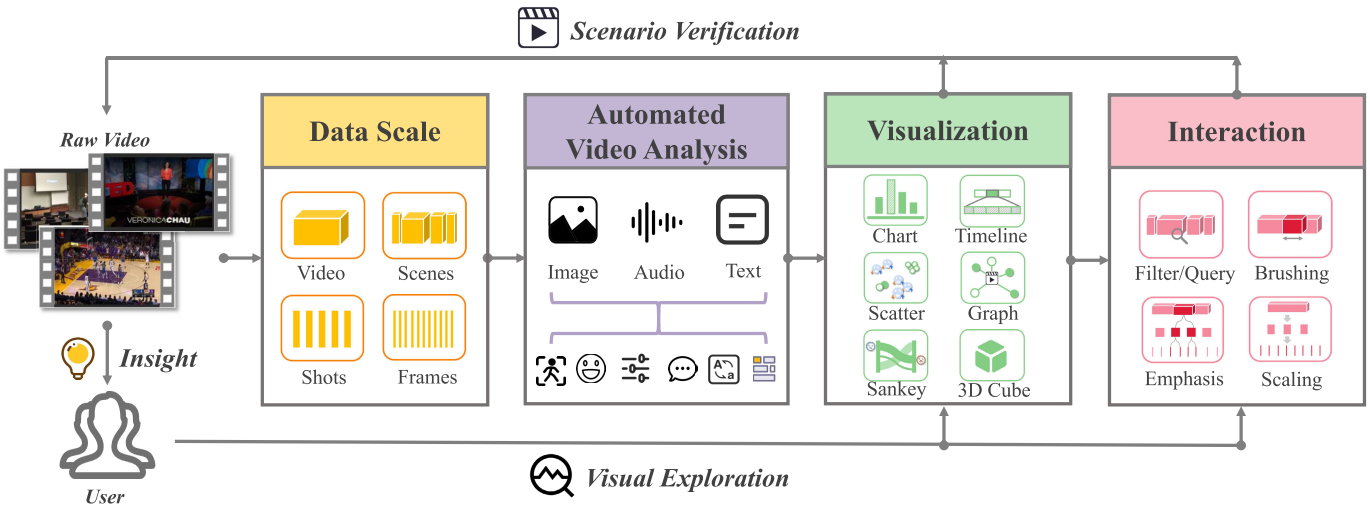



Fig. 1. A general pipeline of video visualization and visual analysis includes these four aspects: (1) **Data scale** consists of *frame*, *shot*, *scene*, *video*. (2) **Automated video analysis**. (3) **Visualization** consists of *chart*, *timeline*, *scatter*, *graph*, *sankey*, *3D cube*. (4) **Interaction** consists of *filter/query*, *brushing*, *emphasis*, *scaling*.

However, these surveys lack awareness of the latest research achievements. In recent years, some video visualization-related surveys have focused on specific analysis tasks [34], [35], [36], [37], application scenarios [38], [39], [40], [41] or interaction techniques [42]. There are also papers that cover a broader range of visualization and visual analysis techniques for both image and video datasets [28]. For example, Afzal et al. [28] conducted a review of visualization techniques for images and videos. Their research primarily focused on exploring the overlaps and differences between computer vision and visualization. They classified articles from various perspectives, including visual interaction, visualization, machine learning methods, data scale, and application domains. During the analysis of the literature, their classification method involved a large number of categories, without providing specific descriptions of the roles of visualization and visual analysis within each category. Compared to this survey, our research focuses on the analysis of visualization and visual analysis techniques. In the analysis process, we employ a multilevel approach to literature analysis, with further subdivisions according to each major category, to enhance the structure of the investigation and provide a clear framework. The multilevel classification framework provides readers with a path from overview to detail in the field of video visualization. This method not only help readers understand and locate various video visualization techniques in this survey, but also conduct comparative analysis between these techniques. Additionally, we discuss in depth how visualization techniques relate to video data in terms of data characteristics, research value, research challenges, and analysis techniques. To the best of our knowledge, there is no state-of-the-art review dedicated specifically to video visualization and visual analysis techniques.

In this survey, we categorize, summarize, and compare state-of-the-art video visualization and visual analysis techniques in multi-dimension, which tend to provide a systematic review. By collecting, filtering, and analyzing relevant papers, we initially provide a description of the design space within the domains of video visualization and visual analysis in Section III. The content of this part can also be seen as a terminology introduction. According to this design space, we develop the description of different research. Subsequently, we organize and categorize existing papers from the perspectives of visual analysis tasks and application scenarios in Section IV and V. Regarding visual analysis tasks, we specifically discuss existing visual analysis techniques in different tasks and elaborate on the limitations and open problems. In the realm of application scenarios, we provide a detailed description of specific scenarios and the corresponding visualization techniques. Finally, we engage in a discussion regarding the prominent and challenging research directions. To facilitate the investigation of our devised taxonomy and reviewed methodologies, we have developed a web-based survey browser. In summary, the contributions of this paper are as follows:

- We collect and summarize 129 typical papers in the research field of video visualization and visual analysis to provide a review.
- We propose a taxonomy from the perspective of analysis tasks and application scenarios on video visualization and visual analysis to offer readers a comprehensive and systematic overview.
- We offer a comprehensive and in-depth examination of the current challenges in this field, and propose potential opportunities for future research.
- We develop a web-based survey browser (https://zjutvis.github.io/VOVideo/) to facilitate the exploration of our taxonomy and associated papers.

## II. Survey Methodology and Taxonomy

### A. Survey Methodology

The objective of this survey is to provide an overview of the existing techniques of video visualization and visual analysis. In order to comprehensively review the existing

| Publications | year | Data Type | | | | Visualization | | | | | | | | Interactions | | | | | | Analysis Tasks | | | | | Evaluation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | image | audio | text | sensor | chart | graph | timeline | projection | glyph | sankey | 3D | filter | scaling | emphasis | editing | brushing | sketch | summarization | understanding | anomaly | augmentation | editing | UWP | VDAR | UP | UE | AP |
| Botchen et al. [88] | 2008 | ■ | | | | | | | | ■ | | | ■ | ■ | | | | | ■ | | | | | | | ■ | | ■ |
| Romero et al. [89] | 2008 | ■ | | | | | | | | ■ | | | | | | | | | ■ | | | | | ■ | ■ | | | |
| Liu et al. [139] | 2009 | ■ | | | | | | | | | ■ | | | | | | | | | | ■ | | | | | | | ■ |
| Zhang et al. [77] | 2010 | ■ | | ■ | | | | | | | | | ■ | | | | | | ■ | | | | | | | ■ | ■ | ■ |
| Adcock et al. [78] | 2010 | ■ | | | | | | | | | | | ■ | | | | | | ■ | | | | | | | ■ | | ■ |
| Chen et al. [153] | 2010 | ■ | | | | | | | | | | | ■ | | | ■ | | | ■ | | ■ | | | | | ■ | | ■ |
| Piringer et al. [54] | 2012 | ■ | | | ■ | | | ■ | | ■ | | | ■ | ■ | | | | | ■ | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| monserrat et al. [81] | 2013 | ■ | | | | | | ■ | | | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Meghdadi et al. [53] | 2013 | ■ | | | | ■ | | ■ | | | ■ | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Hoferlin et al. [86] | 2013 | ■ | | | | | | ■ | | ■ | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Legg et al. [60] | 2013 | ■ | | | | | | ■ | | | ■ | | ■ | | | | | ■ | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Kurzhals et al. [94] | 2013 | ■ | | | | | | ■ | | | | ■ | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Kim et al. [82] | 2013 | ■ | | | | | | ■ | | | | | | | | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Kim et al. [73] | 2014 | ■ | | ■ | | | | ■ | | | | | | | | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Wang et al. [96] | 2014 | ■ | | | ■ | | | ■ | | ■ | | | ■ | ■ | | | | | ■ | | | ■ | | ■ | ■ | ■ | ■ | ■ |
| Dietrich et al. [152] | 2014 | ■ | | | | | | ■ | | ■ | | | ■ | | | | | | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Pavel et al. [76] | 2014 | ■ | | ■ | | | | ■ | | | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Al-Hajria et al. [75] | 2014 | ■ | | | | | | ■ | | | | | ■ | | | | | | ■ | | | | | | ■ | ■ | ■ | ■ |
| Hamid et al. [85] | 2014 | ■ | | | | | | | | ■ | | | ■ | | | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Polk et al. [59] | 2014 | ■ | | | | ■ | ■ | | | | | | ■ | ■ | ■ | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Liao et al. [117] | 2015 | ■ | | | | | | | ■ | | | | ■ | | | | | ■ | | | | | | ■ | ■ | ■ | ■ | ■ |
| Lowe et al. [144] | 2015 | ■ | | | ■ | | | | | | | | | | | | | | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Biswas et al. [79] | 2015 | ■ | | ■ | | | | ■ | | | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Chen et al. [155] | 2015 | ■ | | ■ | ■ | ■ | | ■ | | ■ | | | ■ | ■ | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Duffy et al. [93] | 2015 | ■ | | | | ■ | | | | ■ | | | ■ | ■ | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Jang et al. [52] | 2015 | ■ | | | | | ■ | | | ■ | | | ■ | ■ | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Sun et al. [56] | 2016 | ■ | | ■ | | | ■ | ■ | | | | | ■ | ■ | | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Kurzhalsa et al. [48] | 2016 | ■ | | ■ | | | ■ | ■ | | | | | ■ | ■ | | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Renoust et al. [51] | 2016 | ■ | | | | | ■ | ■ | | | | | ■ | ■ | | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Biresaw et al. [143] | 2016 | ■ | | | | | | ■ | | | | | | | | ■ | | | ■ | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Ma et al. [61] | 2016 | ■ | | ■ | | | ■ | | | | | | | | | | | ■ | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Ma et al. [63] | 2016 | ■ | | ■ | | | ■ | | | | | | | | | | | ■ | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Shi et al. [155] | 2017 | ■ | | ■ | | | ■ | | | | | | ■ | | | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Wu et al. [57] | 2017 | ■ | | | | | | | | ■ | | | ■ | ■ | ■ | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Wu et al. [49] | 2018 | ■ | | ■ | | | | | ■ | ■ | ■ | | ■ | ■ | | ■ | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Wu et al. [148] | 2018 | ■ | | | | ■ | ■ | | | | | | ■ | | | ■ | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Samrose et al. [72] | 2018 | ■ | | ■ | | ■ | ■ | | | | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Chen et al. [102] | 2018 | ■ | | ■ | | ■ | ■ | ■ | ■ | | | | ■ | | | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Halter et al. [154] | 2019 | ■ | | ■ | | | ■ | ■ | | | | | ■ | | | ■ | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| John et al. [55] | 2019 | ■ | | ■ | | ■ | ■ | ■ | | | | | ■ | ■ | | | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| piazentin et al. [119] | 2019 | ■ | | | | | ■ | | | | | | | | | ■ | | | | | | ■ | | ■ | ■ | ■ | ■ | ■ |
| Huber et al. [46] | 2019 | ■ | | ■ | | | ■ | | | | | | | | | ■ | | | | | | ■ | | ■ | ■ | ■ | ■ | ■ |
| Chan et al. [92] | 2019 | ■ | | | ■ | | ■ | | | | | | | | | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| AilieFraser et al. [80] | 2019 | ■ | | ■ | | | ■ | | | | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Fan et al. [115] | 2019 | ■ | ■ | | | | ■ | | | | | | ■ | | | | | | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Andrienko et al. [135] | 2019 | ■ | | | | | | | | ■ | | | ■ | | | | | | | | | ■ | | ■ | ■ | ■ | ■ | ■ |
| Ma et al. [10] | 2020 | ■ | | ■ | | | ■ | ■ | | ■ | | | ■ | | | | | ■ | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Zeng et al. [3] | 2020 | ■ | | | | ■ | | ■ | | ■ | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Guo et al. [108] | 2020 | ■ | | | | ■ | | ■ | ■ | | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Tang et al. [2] | 2021 | ■ | ■ | | | | | ■ | | | | | ■ | ■ | | | | | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Sun et al. [45] | 2021 | ■ | | | | ■ | ■ | ■ | ■ | ■ | | | ■ | ■ | ■ | | | | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Li et al. [146] | 2021 | ■ | | | | ■ | ■ | | | | | | ■ | | | | | ■ | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Ye et al. [147] | 2021 | ■ | | | | ■ | ■ | | | ■ | | | ■ | | | | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Xie et al. [109] | 2021 | ■ | | | | | | ■ | | | | | ■ | ■ | | | | | ■ | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Chen et al. [126] | 2021 | ■ | | | | | | | | ■ | | | | | | ■ | | | | | | ■ | | ■ | ■ | ■ | ■ | ■ |
| Deng et al. [128] | 2021 | ■ | | | | | | ■ | | ■ | | | | | | ■ | | | | | | ■ | | ■ | ■ | ■ | ■ | ■ |
| Soure et al. [98] | 2021 | ■ | ■ | ■ | | | | ■ | | ■ | | | ■ | | | ■ | | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |
| Chu et al. [137] | 2021 | ■ | | ■ | | | | | | ■ | | ■ | ■ | | | | | | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Chung et al. [127] | 2021 | ■ | | ■ | | | | | | ■ | | | ■ | | | | | | | | | | ■ | ■ | ■ | ■ | ■ | ■ |
| Lan et al. [105] | 2021 | ■ | | | | | | ■ | | ■ | | | | | ■ | | ■ | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Wu et al. [133] | 2021 | ■ | | | ■ | | | | | ■ | | | | | | | | ■ | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| zeng et al. [7] | 2022 | ■ | | | | | | ■ | | | | | | ■ | | | ■ | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Liu et al. [27] | 2022 | ■ | | | | | | | | ■ | ■ | | | | | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Chen et al. [6] | 2022 | ■ | | ■ | | | | | | ■ | | | ■ | | | ■ | ■ | | | | | | ■ | ■ | ■ | ■ | ■ | ■ |
| Swift et al. [84] | 2022 | ■ | | | | | | | | | | | | | | | | | | | ■ | | | | ■ | ■ | ■ | ■ |
| Lin et al. [5] | 2022 | ■ | | | | ■ | | | | ■ | | | | | | | ■ | ■ | | | ■ | | | | ■ | ■ | ■ | ■ |
| Seebacher et al. [106] | 2023 | ■ | | | | | | | | ■ | | | | | | ■ | ■ | ■ | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| Chen et al. [4] | 2023 | ■ | | | ■ | | | | | ■ | | | | | | ■ | | | | | | | ■ | | ■ | ■ | ■ | ■ |
| Wong et al. [26] | 2023 | ■ | ■ | ■ | | ■ | | ■ | | | | | ■ | ■ | | | ■ | | | | ■ | | | ■ | ■ | ■ | ■ | ■ |
| Wu et al. [99] | 2023 | ■ | ■ | ■ | | ■ | | ■ | ■ | ■ | | | ■ | ■ | | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ |
| He et al. [100] | 2023 | ■ | | | | ■ | | ■ | ■ | | ■ | | ■ | | ■ | | ■ | | ■ | ■ | | | | ■ | ■ | ■ | ■ | ■ |

Fig. 2. We list the most cited papers each year and the latest paper in each application. We summarize them based on these four dimensions: *data*, *visualization*, *interactions*, and *analysis tasks*. Data type: image-based, audio-based, text-based, sensor-based. Visualization: chart-based, graph-based, timeline-based, projection-based, glyph-based, sankey-based, 3D-based. Interactions: filter, scaling, emphasis, sketch. Analysis tasks: video summarization, video content understanding, video anomaly detection, video augmentation, and video editing. Evaluation: understanding environments and work practices (UWP), visual data analysis and reasoning (VDAR), user performance (UP), user experience (UE), and algorithm performance (AP).

papers in this field, we employed two primary literature retrieval methodologies: search-based and citation-based. The search-based approach was designed to initiate a preliminary exploration of video visualization and analysis techniques. We utilized two queries ("video" AND "visualization"; "video" AND "analysis") with the anticipation of covering all papers related to both video and visualization. Our main sources of papers were highly influential conferences and journals (*Conferences*: IEEE VIS, CVPR, ECCV, ICCV, ACM CHI, EuroVis, and PacificVis. *Journals*: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Image Processing (TIP), IEEE Transactions on Visualization and Computer Graphics (TVCG), IEEE Transactions on Multimedia (TMM), IEEE Computer Graphics and Applications (CG&A), and Computer Graphics Forum (CGF).) The citation-based approach starts with the core techniques in this field and subsequently expands the scope by checking citations and references.

Subsequently, we proceeded to read through the titles of each paper to choose papers that potentially related to video visualization. If a title did not explicitly describe the relevance to our surveyed direction, we further examined abstracts and main text to determine its inclusion in our survey. The inclusion criteria were as follows: (a) the paper involved visualization of video data, and (b) the paper involved exploration and analysis of video data while utilizing visualization techniques. Within the literature collection process, we restricted the time range from 2008 to 2023. Ultimately, we obtained 129 papers specifically focused on video visualization and visual analysis.

Additionally, we conducted a statistical analysis of these papers from dimensions such as journals, conferences, and years (as illustrated in Fig. 3). The statistical information demonstrates the substantial presence of this research direction within top journals and conferences in the visualization domain. Additionally, there has been a continuous increase in research efforts, particularly in recent years. This growth can be attributed to the advancements in deep neural networks, which have propelled the progress of computer vision techniques and provided the basis for the effective implementation of video visualization and visual analysis. At the same time, against the rapid advancements in computer vision, there is an urgent need for visual analysis to serve as a bridge facilitating effective communication between humans and machines.

### B. Taxonomy

To provide an overview of all the papers from various dimensions and generate a more non-mutual classification, we first conducted content summarization and extracted key information from each paper. We collected information including motivation, challenges, contributions, datasets, data processing techniques, target users, visualization techniques, interactive methods, and future work. This information was stored in an Excel table for easy querying and filtering. We then formulated this information into a design space. In particular, we drew on traditional visual analysis workflows and described it as three key components: *data*, *visualization*, and *interactions*. Taking into account the characteristics of video data, we refined each visual analysis component with subcategories (details of the design space are discussed in

Section III). The design space can also be regarded as a terminological explanation of the visual design process, which can provide readers with a structured and comprehensive understanding of these papers. This is also beneficial for researchers who lack a background in video visualization, as it facilitates a comprehensive understanding of the role of each term in the field of video visualization. It enables researchers to more deeply comprehend the principles and methodologies of various techniques throughout the visual design process. During the review of the literature, we utilized selected terms from the design space to summarize existing techniques. Moreover, on our web-based survey browser, researchers can filter and search for articles based on the terms in the design space.

The design space of each paper is determined by the analysis tasks and application domains. We initially classified visual analysis tasks in existing work into four categories based on their appearance frequency: *video summarization*, *video anomaly detection*, *video content understanding*, and *video editing/enhancement*. However, during the classification process, some authors raised concerns about merging video editing and enhancement into a category, as they have significant differences in the context of visualization and visual analysis tasks. After thorough discussions with other authors, we reached a consensus and separated the original category into two distinct categories: *video enhancement* and *video editing*. Additionally, we summarized four common application scenarios: *surveillance*, *sports*, *entertainment*, and *education*.

In addition to describing the papers from the perspectives of visualization workflows, analysis tasks, and application scenarios, we have also summarized the evaluation section of the papers to compare the technique performance. Comparing the completeness of their evaluation can reveal the practical utility in different application scenarios and the capabilities to meet specific user requirements. To enable performance comparison within the same framework, we utilized Isenberg et al.'s [43] adaptation of Lam et al.'s taxonomy [44] to characterize visualization evaluation forms. We adjusted the forms of evaluation according to the area of our survey. Ultimately, four categories were retained: "*Understanding Environments and Work Practices*", "*Visual Data Analysis and Reasoning*", "*User Performance*", "*User Experience*", and "*Algorithm Performance*".

Once the major description dimensions of papers were determined, we proceeded to code all the papers based on these dimensions. Note that the design space is divided into a fine-grained categorization, with each paper classified based on one or more design choices across various dimensions. In contrast, the dimension of analytical tasks or application domains is broader, reflecting the key objectives and research focus of the papers. Therefore, we code each paper with one specific task and scenario to assist readers in understanding the major contribution of each paper, which enhances the readability and practicality of this survey. To ensure the reliability of the coding process, each paper was coded by at least two authors. As a result, three authors independently coded 50 papers and engaged in discussions to resolve any coding ambiguities, aiming to enhance the accuracy and credibility of the coding. To provide a comprehensive presentation, we selected the most cited 73 papers from all 129 papers to display the coding results (Fig. 2). We considered the distribution of papers in dimensions during the selection process and balanced the number of papers in each category. In addition, we also pay attention to the year distribution of these papers to ensure that the papers in different years are covered.

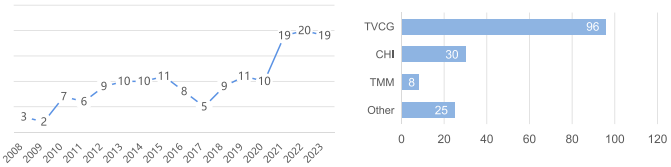



Fig. 3. Left: number of papers over the years. Right: statistics on the number of papers in different journal.

The remaining sections of the paper are organized as follows. In Section III, we introduce the design space used to describe visualization design process. Then, in Section IV, we provide a comprehensive overview of the state-of-the-art techniques for each visual analysis task. In Section V, we present an overview of common application scenarios of video visualization to offer intuitive guidance. Finally, we discuss the challenges and opportunities in this research area. Through this organization, we anticipate that this survey will provide researchers and practitioners with a comprehensive understanding of existing work from multiple perspectives. The structure of our survey is depicted in Fig. 4.

## III. Design Space

In this section, we introduce a design space to describe the characteristics of the visualization design process. As outlined in Section II, the foundation of this design space is structured around three primary dimensions motivated by the visual analysis pipeline: "data", "visualization", and "user interactions". These dimensions encompass the core components of the visualization pipeline. By classifying research papers according to these dimensions, we further refine the design space by identifying and describing commonly utilized subcategories within each dimension. This design space offers a comprehensive and detailed description of research papers, enriching researchers' understanding of video visualization techniques through a visual analysis process-based perspective.

### A. Data Scale

**Frames:** Video is composed of a series of continuous frames, and each frame represents a still image within the video. Frames are the smallest unit of video data. Tang et al. [2] adopted the frame as the fundamental unit for risk assessment of e-commerce videos, facilitating auditors in achieving efficient scrutiny of non-compliant video moderation. Sun et al. [45] utilized the transitions between frames to mine key patterns in abnormal fragments.

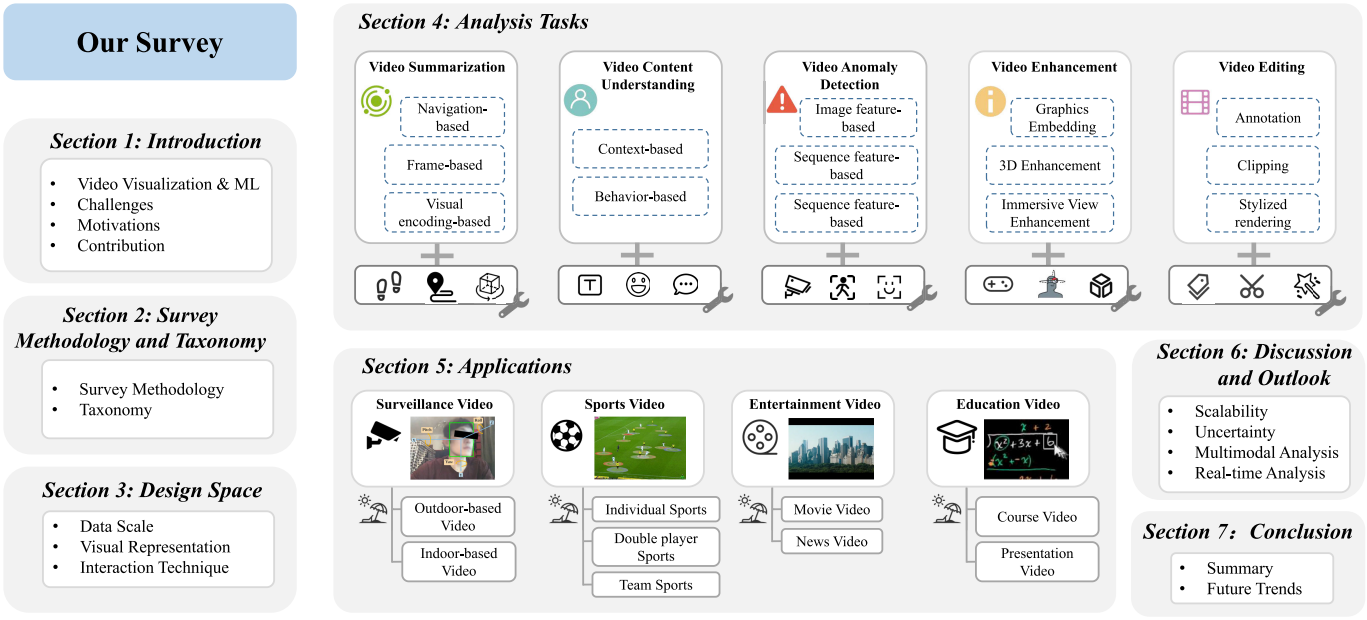



Fig. 4. Overall organization of our survey paper. Section I is the introduction. Section II provides the methodology and taxonomy of our survey. In Section III, we introduce the design space used to describe the visualization design process. Then, in Section IV, we provide a comprehensive overview of the current state and the latest solutions for each visual analysis task. In Section V, we present an overview of common applications of video data to offer more intuitive guidance for practitioners in visualization and visual analysis. Finally, we discuss the challenges and opportunities in Section VI. Section VII is the conclusion.

**Shots:** A shot is a single event that can be considered as a semantic unit, such as an action, a sentence, etc. Shots can be analyzed to mine patterns and generate insights of a particular entity or scene. For example, Huber et al. [46] introduced a lens-oriented video editing tool, serving as a creative aid for aspiring filmmakers to produce videos of remarkable caliber and mesmerizing allure. Truong et al. [47] divided the makeup steps in the video into multiple shots based on the detection of key actions, and conducted hierarchical analysis.

**Scenes:** A scene refers to a collection of shots captured within a specific temporal interval, where the camera's position and perspective remain predominantly unchanged. Kurzhals et al. [48] used scripts to conduct scene segmentation for movies, leveraging multimodal cues to assist viewers in comprehending the plot dynamics and narrative transitions of the movie. Sun et al. [11] divided the scenes based on the Danmu conversation timeline and formed a tree-like visual summary for each scene.

**Video:** A video is organized by multiple scenes, encompassing various shots and an abundance of frames. Extracting and processing features at the video level can be applied to tasks such as video classification, video summarization, video retrieval, and more. For example, Chen et al. [48] extracted and visualized compact summaries of storylines for efficient representation and fast overview of TV programs. Wu and Qu [49] achieved visual analysis of speech video style by converting speech videos into vectors containing multimodal information and projecting them into two-dimensional space.

### B. Visual Representation

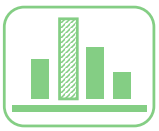
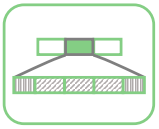
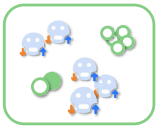


**Chart-based visualizations:** Visualization charts, such as bar charts, pie charts, and stacked charts are often used to show the distribution of video features. For example, Zeng et al. [3] used stacked charts to show the percentage of students' emotions in course instruction. Li et al. [50] employed bar charts to represent the risk levels of cheating among students during online exams.



**Timeline-based visualizations:** The timeline is a visualization form that best matches the characteristic of the video sequence, which presents the abstract information or semantic features extracted from the video in a chronological order. These features are usually represented in terms of color, texture, and shape. For instance, EmotionCues [3] employed multiple colors in a timeline to encode multimodal emotional features, representing the temporal changes in the speaker's emotions within the video. EmoCo [9] employed a multi-level timeline visualization to facilitate the exploration of speech content from the sentence level to the word level.



**Scatter-based visualizations:** Scatter-based visualizations are typically used to project data extracted from videos, where each point can represent a frame or an action. The scatter form can assist the user in understanding the distribution of the video content, such as exploring clusters of frames with similar

features or anomalies. For example, Zeng et al. [7] projected the speech content and gesture content from speech videos separately and uses skeleton-based glyph to represent gestures.

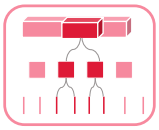
**Graph-based visualizations:** Graph-based visualizations are commonly used to represent relationships between frames, objects, or metadata within videos. They enable users to gain insights and analyze the structure and relationship within the video. For example, Renoust et al. [51] represented the relationships between politicians in news videos using graph-based visualization, assisting users in discovering patterns of political connections among public figures. Jang et al. [52] utilized a node-link diagram to represent the transitions between posture clusters, where each node visualizes a posture using a skeleton-based glyph.

**Sankey-based visualizations:** Sankey-based visualization forms are often used to show the relationship between different modal information or the hierarchical relationship between different fine-grained information in a video. For example, Wu et al. [49] summarized the postures, gestures and rhetoric patterns through sankey diagrams to assist users in learning expressive techniques for public speaking.

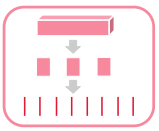
**3D cube-based visualizations:** 3D cube is a form of visualization that restores the two-dimensional spatio-temporal information compressed by the camera into three-dimensional space. This visualization can present the original video content, such as the motion trajectory of the objects in the video, or the change of eye viewpoint when the human watches the video. Meghdadi and Irani [53] summarize the video by presenting the pedestrian trajectory in a 3D cube.

### C. Interaction Technique

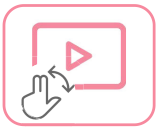
**Filter/Query** interaction technique allows users to query and filter video content of interest based on specific criteria, thereby narrowing down the analysis scope and improving analysis efficiency. Types of filtering [11], [53], [54] include frame filters based on similar or unusual frames, time filters based on video time period selection, event filters based on specific events, attribute filters based on frame attributes or object attributes, and object filters based on moving objects.



**Emphasis** interaction technique enables users to highlight and emphasize specific regions or objects through visual effects like highlighting, zooming, or other visual techniques [9], [55], [56], [57], [58], facilitating the presentation of meaningful patterns. Users can use this approach to highlight key information or content of interest and view further details.



**Scaling** interaction technique enables users to observe content of interest at different levels of granularity, facilitating their exploration of patterns and features within the video from a global overview to local details [50], [54], [59]. Through scaling, users can explore patterns, trends, and correlations within the video progressively.



**Sketch** interaction technique provides users with a natural and intuitive way to create hand-drawn sketches, which can be applied to video annotation, visualization, and analysis [60], [61], [62]. Through sketchs, users are able to record and express their thoughts and opinions on video content in a convenient manner. Ma et al. [63] utilized sketch interaction to enable users to express and capture their ideas more naturally without relying on complex menus or toolbars for annotating keyframes or important segments in videos.

## IV. Visual Analysis Tasks

In this section, we introduce typical analysis tasks in the field of video visualization and discuss the existing visual analysis techniques along with their limitations and open problems for these tasks. We categorize these analysis tasks into five major types: **Video summarization** aims to generate a concise and compact visualization to represent video content. **Video content understanding** focuses on gaining insights from the contextual information of the video. **Video anomaly detection** is employed to identify abnormal patterns in the video stream. **Video augmentation** is intended to enhance the visual effect of the video. **Video editing** is used for narrative annotation and editing of the video. Additionally, we have created a summary table (Fig. 5), which overviews and navigates the visual analysis techniques from the perspective of analysis tasks, encompassing these dimensions: *data processing techniques*, *visualization* and *interaction techniques*, as well as *application scenarios*.

### A. Video Summarization

Video summarization aims to present videos in a concise and compact visual format to convey the key content and main points. By employing techniques such as content extraction, keyframe detection, and temporal compression, the length of the video can be reduced while preserving crucial information [68], [69]. As a result, viewers can quickly grasp the entire video content within a limited timeframe. This technique has been used in many applications such as surveillance [66], [70], entertainment [71], and education [64], [65], [72]. According to the summarization demand in different scenarios, existing techniques can be divided into three types: **navigation-based**, **frame-based**, and **visual encoding-based**.

**Navigation-based technique** is applied in various domains such as online education [73], [74] and audio visual analysis [75], [76], [77], emphasizing efficient navigation and browsing of video data. They aim to assist users in quickly locating the content of interest and provide a more

| Task Description | Data Processing Technique | Visualization Technique | Interaction Technique | Application Scenario | Evaluation |
|---|---|---|---|---|---|
| **Video Summarization** aims to extract key information which can reflect the structure of the video in a compact visualization. | **Navigation-based technology:**<br>• Key frame detection [76], [81], [47]<br>• Text recognition [74], [78]<br>**Frame-based technology:**<br>• Key frame extraction [88], [71]<br>• Object tracking [83], [86] (e.g., detection of pedestrians and athletes)<br>• Multi-View video learning [66], [85]<br>**Visual encoding-based technology:**<br>• Trajectory Recognition [53], [60] (e.g., detection of pedestrians and vehicles)<br>• Semantic information extraction [10], [52], [61] (e.g., sentiment, activity, events, and scenes) | • Graph-based [90]<br>• Timeline-based [53], [74], [76], [81]<br>• 3D cube-based [94]<br>• Map-based [10], [61] | • Scaling [52], [53]<br>• Filter [53], [60], [74], [76], [78], [81], [85], [86], [94]<br>• Sketch [10], [61] | • Surveillance [66], [70]<br>• Entertainment [77]<br>• Educational [64], [65], [73], [78], [82] | • **UWP** (16.7%)<br>• **VDAR** (41.7%)<br>• **UP** (75%)<br>• **UE** (91.7%)<br>• **AP** (50%) |
| **Video Content Understanding** aims to model the video into a state sequence from a specific task, which supports users in conducting hierarchical analysis of video content and promotes further reasoning. | **Context-based technology:**<br>• Topic modeling [55], [95]<br>• Multimodal data analysis [11], [63] (e.g., audio, image, text, and sensor)<br>• Semantic information extraction [154] (e.g., color)<br>**Behavior-based technology:**<br>• Semantic information extraction [7], [8], [27], [101] (e.g., sentiment, activity, events, and scenes)<br>• Multimodal data analysis [9], [49] (e.g., audio, image, text, and sensor)<br>• Frequent itemset mining [96], [101]<br>• Clustering Algorithms [105], [106], [108]<br>• Association analysis [51] | • Chart-based [8], [55], [108], [148]<br>• Graph-based [11], [51], [63], [95], [148]<br>• Timeline-based [9], [11], [51], [49], [55], [96], [101], [106], [154], [105], [108], [148]<br>• Projection-based [9], [49], [55], [101], [108]<br>• Glyph-based [7], [9], [27], [49]<br>• Sankey-based [9], [49]<br>• Map-based [63] | • Filter [7], [8], [9], [11], [51], [49], [55], [96], [101], [106], [105], [106], [108], [148]<br>• Scaling [7]–[9], [11], [51], [49], [55], [96], [109] | • Surveillance [96]<br>• Educational [7], [72], [98] | • **UWP** (80%)<br>• **VDAR** (60%)<br>• **UP** (40%)<br>• **UE** (80%)<br>• **AP** (0%) |
| **Video Anomaly Detection** refers to detecting a video image or object state that does not meet expectations. | **Image recognition technology:**<br>• Image feature extraction [1], [113]<br>• Anomaly Image Detection [1], [113]<br>**Sequential Anomaly detection technology:**<br>• Temporal Transition Pattern Mining [45]<br>• Temporal Event Prediction [54]<br>• Multimodal Event Detection [2], [58] | • Chart-based [26], [45], [50]<br>• Graph-based [45]<br>• Timeline-based [2], [26], [45], [54], [50]<br>• Projection-based [45] | • Scaling [2], [26], [45], [54]<br>• Filter [2], [26], [45], [54], [58], [50]<br>• Emphasis [45]<br>• Sketch [50] | • Surveillance [1], [2], [26], [113], [115] | • **UWP** (40%)<br>• **VDAR** (60%)<br>• **UP** (60%)<br>• **UE** (80%)<br>• **AP** (60%) |
| **Video Editing** entails the storytelling annotation , clipping or stylized rendering of original videos to enhance their comprehensibility or achieve specific narrative effects. | **Video Annotation technology:**<br>• Event Detection [118], [119], [128]<br>• Object Tracking [4], [6], [128], [126] (e.g., tracking the basketball/table tennis and athlete)<br>• Pose estimation [4], [6], [128], [126] (e.g., detecting the athletes poses during sports competitions)<br>**Video Clipping technology:**<br>• Recommendation System [46]<br>• Shot Boundary Detection [122]<br>• Image Segmentation [122]<br>**Stylized rendering technology:**<br>• Image style transfer [123], [129]<br>• Video synthesis [62], [124] | • Timeline-based [119], [128]<br>• glyph [4], [6], [128], [126] | • Filter [6]<br>• Sketch [62] | • Sports [4], [6], [119], [128], [126]<br>• Entertainment [120]<br>• Educational [122] | • **UWP** (14.3%)<br>• **VDAR** (42.9%)<br>• **UP** (60%)<br>• **UE** (80%)<br>• **AP** (60%) |
| **Video Enhancement** employs the means of visualization techniques to elevate the quality and viewing experience of videos. | **Graphical embedding:**<br>• Key frame extraction [153]<br>• Sequence Modeling [5], [153]<br>• Clustering Algorithms [5]<br>• Action Recognition [5], [135]<br>• Trajectory Recognition [5], [135], [136], [131]<br>**3D augmentation:**<br>• Depth-based 3D Reconstruction [138]<br>**Immersive perspective enhancement:**<br>• Viewpoint matching algorithms [139], [140] | • Chart-based [5]<br>• Glyph-based [5], [135], [136], [131] | • Filter [5], [135], [153]<br>• Sketch [5] | • Surveillance [131]<br>• Sports [5], [133]<br>• Entertainment [134] | • **UWP** (0%)<br>• **VDAR** (40%)<br>• **UP** (60%)<br>• **UE** (100%)<br>• **AP** (40%) |

Fig. 5. Summary of the state-of-the-art papers from the perspective of analysis tasks, encompassing *data processing techniques*, *visualization*, *interaction techniques*, *application scenarios*, and *evaluation*. Evaluation: understanding environments and work practices (UWP), visual data analysis and reasoning (VDAR), user performance (UP), user experience (UE), and algorithm performance (AP).

user-friendly interactive browsing experience. Based on different navigation approaches, navigation techniques can be further categorized as *query-based*, *progress bar-based*, and *hierarchical-based*.

*Query-based technique* refers to the method of searching and locating specific content in videos using various forms of queries, such as text queries [78], [79], [80], voice queries [64], and image queries [81]. TalkMiner [78] utilizes slide detection algorithms to create searchable text indexes, enabling users to easily search and browse lecture webcasts. Additionally, relying only on existing key information for queries sometimes leads to difficulties in accurately expressing search demands. Therefore, NoteVideo [81] employs an approach to extract geometric shapes, formulas, and other concepts from blackboard-style educational videos. They generate a summary image representing these mathematical concepts, which serves as a navigation interface, enabling users to directly navigate to specific video frames associated with particular concepts.

*Progress bar-based technique* provides users with a comprehensive overview, eliminating the need for full-text searches within videos. Current techniques commonly leverages the segmented video within the progress bar to assist users in nonlinearly locating and navigating to specific segments. By displaying pertinent information such as video keywords [74], thumbnails [65], [73], [82] on the progress bar of the video player, users can gain an intuitive understanding of the overall structure and content distribution of the video. Users can swiftly navigate to areas of interest for playback by either dragging the slider on the progress bar or directly clicking on the thumbnails. Moreover, the progress bar-based navigation also offers specific fast-forwarding techniques for

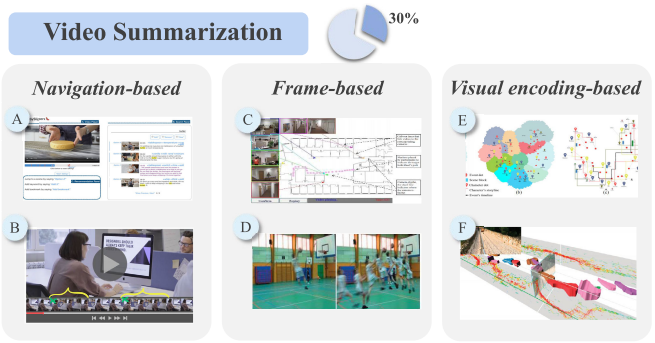



Fig. 6. (A) RubySlippers [64] supports content-based navigation, which can help users browse guide videos more easily (B) Subramanian et al. [65] use visualization and interactive methods to help users understand audiovisual data and gain valuable insights. (C) Wang et al. [66] discuss video placement, spatial context rendering methods, and their impact on path reconstruction tasks. (D) Sunkavalli et al. [67] generate high-quality still images from short video clips. (E) Ma et al. [61] summarize the multimodal movie data in a map. (F) Liang et al. [56] summarize the eye-tracking information and image saliency in a spatial-temporal cube.

controlling the video's playback progress, such as frame skipping [70] and temporal blending [76].

*Hierarchical-based technique* refers to the process of dividing a video into multiple hierarchical structures and providing summaries and thumbnails for each level, enabling users to effectively utilize the hierarchical information to quickly locate segments of interest [47], [76]. This approach is commonly employed for lengthy videos such as courses [47] and documentaries [76]. Existing research typically employs automated processing techniques to convert videos into text and generate segmentation or summaries. For instance, [47] has demonstrated automatic recognition and segmentation of detailed action steps in makeup tutorial videos, with rough categorization based on facial features. However, automated segmentation and summarization methods may not fully satisfy users' flexible requirements. Therefore, Video Digests [76] allows users to manually mark chapters or sections within videos to enhance the accuracy of generated summaries, building upon automated segmentation techniques.

**Frame-based technique** is extensively employed in various domains, including surveillance [66], [83], sports [84], and transportation [85]. These methods consider videos as a series of static images, with each image representing a specific moment in the video. By processing and integrating the information extracted from video frames, they generate comprehensive video summaries that enable users to quickly and intuitively comprehend the themes and content of the videos from a global perspective. Depending on the input data format, these methods can be categorized into two types: *multicamera-based* [84], [86], [87] and *singlecamera-based* [67], [71], [85].

*Multicamera-based* technique fuses and visualizes contextual information from multiple cameras, providing users with a holistic view. One of the main challenges involves integrating video streams from different cameras while maintaining scene consistency and coherence. To address this challenge, existing research first aligns, calibrates, and synchronizes the image data from different cameras, and further fuses them to generate panoramic images or wide-angle videos. Computer vision techniques employed in the fusion process include multi-view video synthesis [87], path reconstruction [66], and viewpoint transformation [84]. The fused videos can be presented using multi-field visualization methods [66], [84] and spatial layout-based visualization techniques [86].

*Singlecamera-based technique* focuses on summarizing the specific semantic content of the videos. They can be further categorized into two subtypes: video key content extraction [67], [71] and spatiotemporal information analysis [85], [88]. Video key content extraction relies on the data mining techniques, such as visual saliency of video frames [67], and semantic information within the video [71]. Furthermore, summarizing spatiotemporal information based on moving objects in videos facilitates a better understanding of object trajectories, interactions, and relationships. Such object information can be presented and analyzed using visualization techniques like trajectory presentation [85] and spatiotemporal relationship graphs [88].

**Visual encoding-based techniques** aims to abstract events, object features, or other metadata in the video as graphic-based symbols, enabling the revelation of patterns within the video. Existing research employs various visualization methods to abstract summarization of video content. These methods primarily include *temporal-based* [72], *2D spatial-based* [89], [90], and *3D cube-based* [53], [91] visualization.

*Temporal-based technology* [72], [92] primarily employs timeline-based visual forms to organize and present abstracted object features or events from the video in chronological order, showcasing their temporal sequence and evolution.

*2D spatial-based technique* is no longer confined to the temporal attributes of videos. Instead, it enables users to engage in nonlinear exploration [10], [61], [89]. These techniques extract and abstract the spatiotemporal correlations among events or objects within videos and depict state transitions as structured attribute information. In visualization, hierarchical visualizations [61], map metaphors [10], [61], and other visual forms are employed to provide an overview of the contextual information in videos. Moreover, to present the specific content of videos intuitively, certain works [93] segment videos based on their semantic information and summarize the spatiotemporal variations of objects within video segments through the generation of static spatiotemporal snapshots of moving objects.

*3D cube-based technique* combines temporal information and spatial contextual information from videos [53], [91], [94]. In a 3D cube, one dimension represents temporal information, while the other two dimensions depict 2D-based spatial information. For instance, Liang et al. [94] (Fig. 6(A)) devised a static spatiotemporal cube visualization method to summarize the spatiotemporal distribution of eye-tracking data and image saliency. Although this 3D visualization approach integrates more comprehensive information into a single space, its extensive representation within a 3D volume may impose a significant cognitive load and incur high rendering costs.

*Limitations and Open Problems:* Although existing video summarization techniques have made significant progress, there are still some limitations and potential directions for future research. First, navigation-based techniques show

| Techniques / Metrics | Navigation-Based | Frame-Based | Visual Encoding-Based |
|---|---|---|---|
| **Scalability** | High scalability for lengthy videos on content locating | Low scalability and high computational complexity for lengthy videos | High scalability for abstracting lengthy and multimodal videos |
| **Summarization Quality** | Constrained by predefined rules and model performance | Constrained by model construction of scene consistency and completeness | Constrained by pattern mining techniques and visual design effectiveness |
| **Advantages** | Support efficient content retrieval and large-scale navigation | Correlate video summary with raw video information and provide an integrated view | Increase data-ink ratio and reduce user perception load |
| **Weakness** | Require prior knowledge of specific scenarios | Lack of detailed information on the raw video | Demand considerable learning costs for visual encoding understanding |

Fig. 7. A comparison of three video summarization techniques (*navigation-base*, *frame-Based*, and *visual encoding-based* techniques) on four dimensions: *scalability*, *summarization quality*, *advantages*, *weakness*.

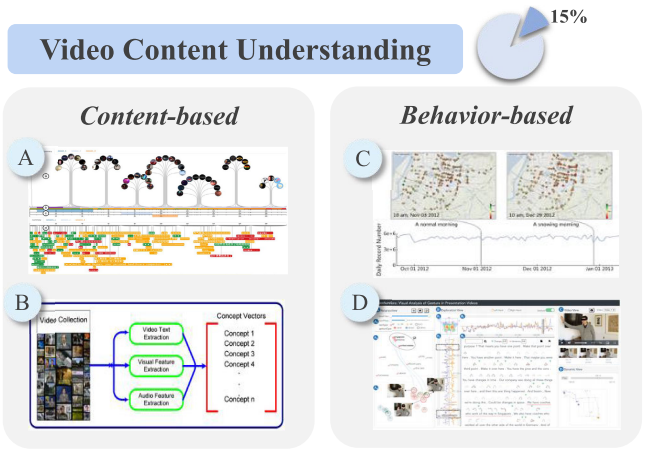



Fig. 8. (A) VideoForest [11] summarizes the movie scenes in a forest-style visualization. (B) This paper [95] builds video data concept vectors and performs semantic similarity calculations to achieve semantic classification. (C) Wang et al. [96] utilized clustering analysis, frequent pattern mining, and anomaly detection to assist users in identifying traffic flow patterns. (D) GestureLens [7] aims to help professional presentation coaches improve their gesture training by analyzing presentation videos.

limitations when dealing with complex and vague queries. Thus, how to transform users' vague intentions of video queries into tasks that are understandable by machines remains an open issue. Second, frame-based techniques provide an intuitive overview of video content, but they fall short in integrating audio, text, and other modalities. This lack of integration fails to meet the need for a comprehensive and in-depth video summarization. Finally, visual encoding-based techniques, while providing key patterns and contextual information of videos, still face challenges in dealing with long-duration videos. These challenges include how to visualize large volumes of video content to generate compact representations, and how to represent them appropriately in detail.

### *B. Video Content Understanding*

Video content understanding involves conducting in-depth analysis and interpretation of the video, including semantic information and contextual relationships. This requires the utilization of computer vision and machine learning techniques, such as object recognition [16], action analysis [3], [27], [57], and scene understanding [97], to extract relevant information from the video. Visual analysis techniques are then applied to gain valuable insights and enhance the understanding of video content. This analysis method has been used in surveillance [96], education [7], [72], [98] and other fields [99], [100]. Depending on the specific focus and scope of content understanding, visualization techniques can be categorized into two categories: **context-based** and **behavior-based**.

**Context-based** comprehension of video content refers to the analysis and comprehension of the overall content of a video, including aspects such as scenes, backgrounds, and plots. This approach involves modeling the spatial and temporal context of the video to identify scenes, understand the plot, and infer the progression of events. Existing visualization methods analyze videos from two perspectives: the multi-level information analysis and the horizontal analysis of multimodal data. The multi-level information, such as scenes, shots, and actions, aiding users in better understanding the structure and information of the video content. The core idea revolves a overview-to-detail exploration, which enables users to selectively focus on and analyze specific video segments of interest while maintaining an understanding of the entire video content. Some studies employ interactive operations, including scaling, panning, and filtering, through scalable timelines [9], [48], [101] to allow users to adjust their focus and level of analysis. To reveal the interconnections between different levels of video content, some research employs graph structures [51] and personalized visual metaphors [63], to enhance the understanding of video semantic structure.

Although existing visualization techniques can help users better explore the content of videos, these techniques only reveal a single modality in the video data. However, videos typically encompass various modalities of information [8], [55], [98], including text, images, and audio. Different modalities complement and validate each other, and the integration and analysis of multiple modalities yield richer contextual semantic information. The challenge in video content understanding lies in the distinct representations and features of different modalities. To address this challenge, existing visualization techniques [49], [101] unifies information from multiple modalities in the video onto shared attributes, such as emotions, postures, and actions, to facilitate communication across modalities. In terms of visualization, researchers have employed Sankey diagrams [9], [49], multidimensional temporal graphs [101], and basic charts [102], [103] to present multimodal attribute information and the interrelationships between modalities. Interactive operations such as filtering and brushing allow users to explore the differences and consistencies between modalities. Furthermore, different modalities may contain redundant information. To integrate meaningful information from different modalities, some studies utilize sequence analysis techniques [9] to filter out key information from each modality.

**Behavior analysis** refers to understanding video content by analyzing moving objects in the video. This approach involves computer vision techniques such as object detection and tracking techniques to extract moving objects from video and analyze their behavior with visual analysis techniques.

Existing visual analysis techniques employ automated algorithms to extract object behaviors and employ pattern mining techniques to discover patterns and regularities in object

| Techniques / Metrics | Content-Based | Behavior-Based |
|---|---|---|
| **Scalability** | High scalability for lengthy videos on multi-level and multi-modal content understanding | High scalability for lengthy videos based on automatic model and pattern mining techniques |
| **Conprehension Ability** | Possess strong analytical capabilities for static information, with rich contextual information | Possess strong analytical capabilities for dynamic information, with rich contextual information |
| **Advantages** | Support hierarchical analysis from overview to detail, and integrated analysis based on multimodal information | Support analysis of dynamic behaviors in videos, including the event pattern mining, trend prediction, and intention understanding |
| **Weakness** | Overlook the dynamic changes in the temporal dimension | Require prior knowledge of specific scenarios |

Fig. 9. A comparison of three video content understanding techniques (*content-based* and *behavior-based* techniques) on four dimensions: *scalability*, *conprehension ability*, *advantages*, *weakness*.

motion. These patterns are then presented and explained through visualizations. Commonly used pattern discovery techniques include clustering analysis [104], [105], [106], [107], frequent pattern mining [96], topic extraction [55], comparative analysis [27], and other machine learning methods [108], which are used to uncover spatiotemporal variations of moving object within the video. For instance, Wang et al. [96] utilized clustering analysis, frequent pattern mining, and anomaly detection to assist users in identifying traffic flow patterns and trends in traffic videos. In a map-based visualization approach, they showcased the trend of traffic patterns over time using a timeline and represented traffic density information with heatmaps. As automatically extracted patterns may not necessarily align with user interests, this research also supported user customization of queries and filtering. Similarly, studies [27], [104], [108], [109] employed timeline-based visualizations to depict the spatiotemporal evolution of video object patterns, while another study [109] provides a visual summary of object motion patterns within a specific video segment using heatmaps.

*Limitations and Open Problems:* Current video content understanding techniques commonly lack interpretability, prompting future research to focus on developing more interpretable methods. The interpretable techniques would help users gain a deeper understanding of the system's reasoning processes while supporting user feedback on model results. In addition, processing long-duration video content continues to be a challenge. Effective modeling and structuring of video content for quick browsing and in-depth analysis still require further exploration.

### C. Video Anomaly Detection

Video anomaly detection aims to identify aberrant patterns and behaviors within video streams [110], [111], [112]. Anomaly is highly correlated with abnormal periods and abnormal subsequences in videos which tends to occupy only a small fragment of the overall video stream. By modeling exceptional behaviors in the videos, it becomes possible to detect non-standard, sudden, or irregular events. This technique finds extensive application in security surveillance [1], [26], [113], enhancing the efficiency and accuracy of video monitoring systems. The existing techniques primarily contain advanced **image recognition techniques** [1], [113], [114] and **sequential anomaly detection algorithms** [2], [26], [54] to detect and recognize complex abnormal behaviors and events in video image sequences.

In terms of the image recognition techniques, some works model the anomaly based on the image features [1], [45], [113]. RipViz [1] integrates machine learning with flow analysis feature detection to extract rip currents from static videos. Utilizing optical flow technology, it captures unstable 2D vector fields from these videos, which aims to analyze the motion of each pixel over time. The identified rip current locations are then overlaid on the original video for an intuitive visualization. Furthermore, some works leverage multimodal anomaly detection algorithms to discover video anomalies [2], [26], [58]. The reciprocal verification of information between different modalities yields more precise results in anomaly detection and analysis. These methods employ deep learning models, such as multimodal convolutional neural networks [115], to jointly analyze multimodal data, achieving enhanced accuracy in detecting and analyzing anomalies.

In terms of visualization, these techniques employ charts [115], custom glyphs [54], and maps [58] to present multimodal data, assisting users in intuitively comprehending the anomalous information within. For example, Tang et al. [2] (Fig. 11(C)) introduced a risk-aware framework named VideoModerator, designed for rapid detection and removal of inappropriate or explicit content in e-commerce live streaming. This work follows a "learn and moderate" strategy, which supports interactive iterative labeling of authentic multimodal video tags, bridging the gap between human moderators and machine learning models.

However, these models for video pattern recognition are a "black box" for users. To solve this problem, recent studies [116], [117] have introduced interpretation techniques in visual predictive analytics to explain these predictive models. Additionally, some works adopt the hierarchical-based approach to scrutinize anomaly information at different levels of granularity, verify the reliability of model results, and capture the complex details of anomalies. For instance, Piringer et al. [54] conducted surveillance on video streams within road tunnels, establishing a context-sensitive priority concept for anomalous information to delineate scenarios ranging from routine operations to catastrophic disasters. The visualization interface presents event information at various levels of abstraction, including time-based abstract event listings, spatial-temporal anomalous events within the tunnel, and detailed monitoring videos corresponding to each event. Similarly, Li et al. [50] detected and analyzed students' head and mouse movements during exams, visualizing suspected cheating behaviors across different hierarchical levels, and facilitating swift identification by educators. This primarily encompasses four levels: a student list view for a quick overview, a problem list view incorporating risk indicators, a behavior view encompassing mouse and head movements, and a replay view showcasing the original video.

*Limitations and open problems:* Although existing research has predefined anomalies based on different contexts, the definition of anomalies often relies on individual experience and specific backgrounds. As a result, there is still a lack of consensus between human intent and model understanding. Future research can leverage visualization techniques to incorporate individual intuition and expertise into the video

| Metrics \ Techniques | Image Feature-Based | Sequence Feature-Based | Multimodal Feature-Based |
|---|---|---|---|
| **Advantages** | Supported by mature techniques and has widespread applications | Support the identification of anomalous patterns in dynamic sequences | Support the analysis of multimodal data, with high reliability of detection results |
| **Weakness** | Overlook anomalies based on time sequence | Require a large amount of sequence data to ensure the detection accuracy | Demand high computational resources |

Fig. 10. A comparison of three video anomaly detection techniques (*image feature- based*, *sequence feature-based*, and *multimodal feature-based* techniques) on two dimensions: *advantages*, *weakness*.

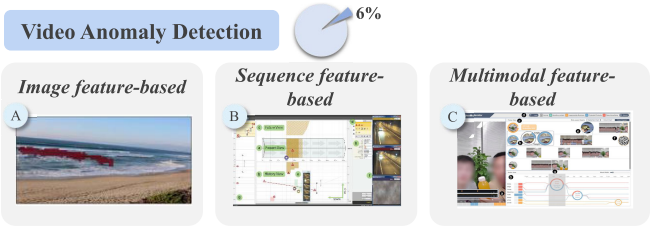



Fig. 11. (A) Piringer et al. [54] conducted surveillance on video streams within road tunnels, (B) Silva et al. [1] introduced a feature detection method based on deep learning and flow analysis, which is used to extract the location of tearing flow from static video and display it visually. (C) Videomoderator [2] presents the multimodal user interface for video moderation.

analysis process, thereby facilitating consensus on anomalies between humans and machines. Additionally, some studies have attempted to utilize multimodal information for anomaly detection, but the interpretability of the model's results is typically lacking. Therefore, future research can explore the interrelationships among multimodal information to assist users in validating the consistency between different modalities, and then better support decision-making.

### D. Video Editing

Video editing entails the **annotation** [118], [119], [120], **clipping** [46], [121], [122], or **stylized rendering** [62], [123], [124], [125] of original videos to enhance their comprehensibility or achieve specific narrative effects. These techniques are mainly used in sports videos [4], [6], entertainment videos [120], and educational videos [122].

**Annotation-based** video editing focuses on annotating objects, behaviors, events, and other elements within videos to generate more understandable video content. This task presents a notable challenge as it involves manual annotation, demanding substantial time and effort. To address this predicament, existing methods primarily alleviate the burden of human annotation through two approaches: *automated video semantic extraction* and *user-friendly interactive video annotation*.

*Automated video semantic extraction techniques* employ advanced algorithms encompassing computer vision, natural language processing, and semantic understanding, to initially identify and extract the information that needs to be annotated. These algorithms facilitate the automatic recognition and annotation of content and objects within videos, thus alleviating human cognitive loads associated with comprehending basic video elements. Subsequently, these techniques also enable users to interactively annotate videos, catering to their personalized annotation requirements. Existing studies [6], [119], [126] usually establish appropriate annotation guidelines based on specific scenarios. Their guidelines encompass the objectives, content, and standards for annotation, as well as the visual representation of annotations. For rules of specific contexts, these guidelines offer annotation recommendations while ensuring the accuracy and consistency of annotations, thereby enhancing the efficiency and reliability of the annotation process. The visual presentation of annotations typically involves embedding scene-specific symbols into the video content, augmenting the efficiency of video comprehension.

To enrich the information conveyed by videos and fulfill various video comprehension demands, certain studies [6] introduce supplementary information such as textual comments, viewer perspectives [4], and question-answer pairs [127] as annotations for videos. Furthermore, in addition to enhancing video comprehension through visual annotations, some works [128] employ these annotated data to support the development of video event recognition models.

*User-friendly interactive techniques* [120], [128] emphasize the design of user-friendly interfaces and intuitive interaction methods, facilitating user convenience in video editing and annotation processes. For instance, EventAnchor [128] (Fig. 13(A)) is a user-centric annotation tool composed of a selector and an annotator. This annotator enables users to choose event types from a predefined list and apply them to specific time intervals within the video. Furthermore, once users have selected one or multiple event types, they can utilize the annotator to append further context-specific details.

**Clipping-based** video editing focuses on the cutting and arrangement of video clips to create a specific narrative. Existing techniques employ automated algorithms to identify keyframes, segments, or events, assisting users in quickly marking significant moments. This allows users to concentrate on the creative process itself rather than the editing process, thereby enhancing video editing efficiency. In addition to editing raw video materials, editors can enhance the artistic appeal and attractiveness of videos by incorporating additional elements. These supplementary elements can include images, audio, animations, etc. In terms of visualization, certain works [46], [122] provide a key-information-based timeline and thumbnails for videos, enabling editors to better comprehend the structural content, temporal relationships, and logical connections between segments. This method is convenient for users to precisely control the time duration, and transition of the video in the process of editing the video.

Although timeline-based editors support intuitive and flexible editing, these tools may only be suitable for individuals with some video editing experience, and require a learning curve for novice creators. To improve editing efficiency for novice creators, several works [46], [122] summarize commonly used editing techniques and offer recommendations and guidance (such as editing skills, transition techniques, audio processing methods, subtitle design techniques, etc.) to editors. For instance, B-Script [46] (Fig. 13(B)) leverages data-driven recommendation algorithms to provide valuable references for editors, aiding them in making decisions about adding supplementary information.

**Stylized video rendering** is to enhance the visual presentation and viewing experience of videos by applying artistic effects and style transformation techniques. Existing

| Techniques / Metrics | Annotation | Clipping | Stylized rendering |
|---|---|---|---|
| **Scalability** | High scalability for annotating key content in lengthy videos | Low scalability for constructing the narrative of lengthy videos | High scalability and high computational complexity for lengthy videos |
| **Editing Quality** | Constrained by predefined rules and model efficacy in semantic extraction | Constrained by recommendation efficacy and model efficacy | Constrained by the consistency of model construction in object details and appearance |
| **Advantages** | Support efficient key content locating and large-scale annotation | Support flexible and creative narrative construction | Increase the artistic effect and visual impact of the video |
| **Weakness** | Require prior knowledge of specific scenarios | Require video editing experience and the editing process is time-consuming | Demand high computational resources and may cause distortion from the raw video |

Fig. 12. A comparison of three video editing techniques (*annotation*, *clipping*, and *stylized rendering* techniques) on four dimensions: *scalability*, *editing quality*, *advantages*, *weakness*.

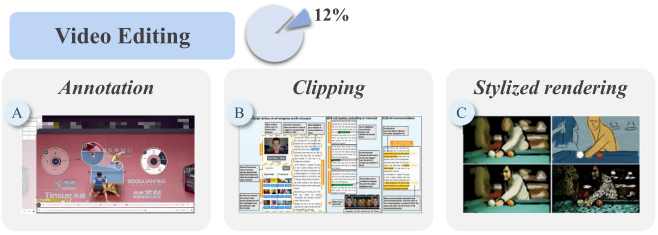



Fig. 13. (A) EventAnchor [128] is a user-centric annotation tool composed of a selector and an annotator. (B) B-Script [46] leverages data-driven recommendation algorithms to provide valuable references for editors. (C) AniPaint [123] supports interactive control over individual regions and keyframe settings for individual strokes by users.

works predominantly employ machine learning and artificial intelligence techniques such as image style transfer [123], [129], [130] and video synthesis [62], [124] to achieve artistic rendering of videos. The difficulty in implementing these techniques is to maintain the consistency of video object details and appearance. Specifically, it provides a unified abstraction and temporal coherence while accurately rendering the art style in terms of details and object boundaries. Furthermore, to address the need for fine-grained and personalized rendering, some studies combine automated painting rendering with detailed interactive control, utilizing either stroke-based approaches [123] or gesture-based approaches [62].

**Limitations and Open Problems:** Existing video editing techniques primarily focus on two aspects: video editing skills [46], [122], [123] and automated semantic extraction techniques [4], [6], [128]. However, these studies are not sufficient in mining and understanding users' editing intentions. Accurately interpreting and expanding users' editing intentions, and subsequently providing editing suggestions that align with their needs, can significantly reduce the editing burden on users. This is especially important for novice users, as their editing intentions tend to be more ambiguous. Furthermore, with the rapid growth of video content, the demand for video editing has also increased. However, existing techniques do not support batch video editing. There are several challenges in this process: (1) the effective presentation of batch videos. (2) Understanding users' editing intentions and extending these intentions to batch videos. (3) Validating that batch editing operations align with users' intentions. These are open problems that have not been fully explored.

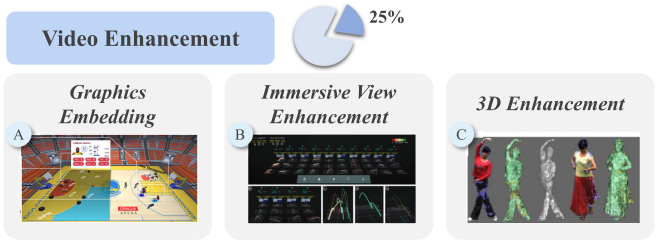



Fig. 14. (A) Lin et al. [5] describes a technique for enhancing the viewing experience of basketball games through embedded visualization technique. (B) TIVEE [137] is an immersive visual analysis system designed to help users explore and explain badminton tactics from multiple perspectives. (C) Liu et al. [139] introduced a point cloud-based multi-view stereo matching algorithm for free-viewpoint video.

## *E. Video Enhancement*

Video enhancement techniques employ visualization techniques to enhance the video quality and viewing experience, including enhancing video content, improving visual effects, and improving viewers' understanding of videos. These techniques have been used in many fields such as surveillance [131], [132], sports [5], [133], and entertainment [134]. Based on different technical characteristics, current techniques can be mainly divided into three types: **graphical embedding-based** [5], [132], [135], [136], **immersive-based** [137], [138], and **3D augmentation-based** [139], [140].

**Graphical embedding-based techniques** involve the integration of visual elements, such as graphics [126], [141], [142], labels [143] and charts [141], into raw videos to provide annotations and contextual cues. By inserting visual representations such as arrows, lines, shapes, crucial points, and actions into the video, the video content can be clarified and accentuated, thereby enhancing the viewer's understanding and attention. However, unsuitable embedding may hinder users from gaining insights into the video data. To tackle this challenge, numerous rule-based visualization methods have been proposed. For instance, Stein et al. [136] summarize the game movement states in soccer videos and design corresponding bespoke visual forms (e.g., heat maps, doughnut charts, and timeline movement paths). VisCommentator [126] summarizes objects, events, and tactics in table tennis matches and recommends contextualized visual forms to users. It enables personalized adjustments of the semantic information embedded in the video annotations, facilitating the production of high-quality augmented videos.

**Immersive-based techniques** utilize immersive interactive devices, such as head-mounted displays [134], [144], [145], to enhance the viewpoints of videos, thereby heightening the viewer's sense of engagement and immersion. This technique provides diverse perspective choices, such as third-person and first-person viewpoints, enabling viewers to observe video content from varying angles and obtain distinct insights and experiences. However, realizing such immersive interaction also confronts several key challenges: the avoidance of visible distortions and discontinuities at depth discontinuities. To overcome these issues, Ana et al. [138] adopt an image-based rendering approach that converts depth information into a 3D grid and perform corresponding translations

| Techniques / Metrics | Graphics Embedding-based | Immersive-based | 3D Enhancement-based |
|---|---|---|---|
| **Enhancement Effect** | Constrained by predefined rules and model performance | Constrained by high-performance hardware support and precise processing of depth information | Constrained by model construction of visual consistency and coherence |
| **Advantages** | Support precise visual guidance for key content in videos | Support immersive and multi-perspective video viewing experiences | Supporting realistic and stereoscopic video visual effects |
| **Weakness** | Demand considerable learning costs for visual encoding understanding | Lack enhancement of detailed and context information in video | Demand high computational complexity |

Fig. 15. A comparison of three video enhancement techniques (*Graphics Embedding-based techniques*, *Immersive-based techniques*, and *3D Enhancement-based techniques*) on three dimensions: *enhancement effect*, *advantages*, *weakness*.

based on viewers' head movements to generate new perspectives.

**3D augmentation-based techniques** offer viewers a more realistic viewing experience. By transforming two-dimensional videos into three-dimensional format, viewers can access videos with unrestricted viewpoints, perceive the depth and distance of objects within the video, and acquire a heightened sense of authenticity. Existing methods [139], [140] address the data sparsity and discontinuity caused by camera position changes by proposing viewpoint-matching algorithms to ensure visual consistency and coherency.

*Limitations and Open Problems* Current annotation-based techniques primarily focus on enhancing specific frames, actions, and shots, rather than a comprehensive understanding of the overall video content. While automated detection and annotation recommendations have improved the convenience of user interaction, video annotation remains a cognitively demanding task for users who are not familiar with the video content. In future work, annotation enhancement of video content can be developed based on concepts or questions proposed by users, further enhancing the accessibility and understanding of video content. Secondly, although existing immersive-based techniques provide users with a fully immersive interactive experience, most methods enhance video content only through the visual channel, ignoring other sensory channels such as hearing and touch. Moreover, research on user experience evaluation (such as the authenticity of user perception, the naturalness of interaction, and the comfort of long-term use) for immersive-based video enhancement techniques is still insufficient. Lastly, 3D augmentation-based techniques, due to their high computational complexity, are difficult to apply in real-time video analysis and have certain limitations in terms of user interaction.

## V. Applications

In this section, we categorize video data according to application scenarios and summarize four major types of video data: **surveillance video** for indoor scene and outdoor scenes, **sports video** for individual sports, double-player sports and team sports, **entertainment video** for movies and news, **education video** for course video and presentation videos. For each application scenario, we first describe the characteristics of the video data, then discuss the research value, challenges, and the existing visual analysis techniques. In discussing visual analysis techniques, we also provide a discussion based on the visual analysis tasks summarized in Section IV.

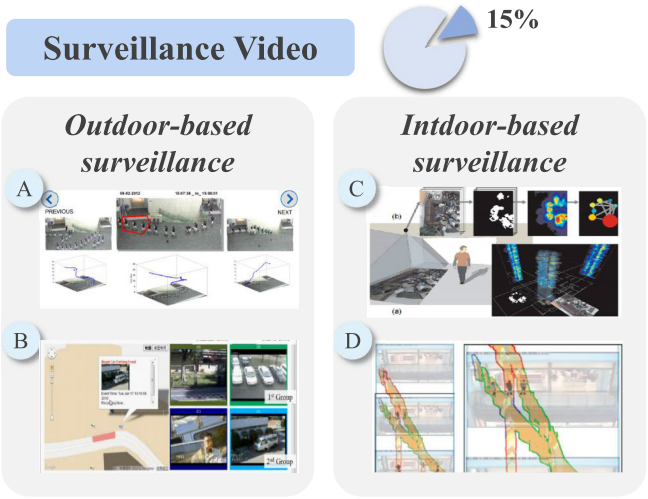



Fig. 16. (A) sViSIT [53] maps the spatial-temporal trajectory into a 3D cube. (B) Fan et al. [58] utilizes IBM-S3 for temporal anomaly detection and employ a matrix arrangement visualization approach. (C) Viz-A-Vis [89] maps the activity space into a 3D cube for visual analysis and exploration. (D) Botchen et al. [88] propose the video stream summarization technique based on multi-attribute mapping.

Additionally, we have constructed a summary table (Fig. 17), which overviews and navigates the visual analysis techniques from the perspective of application scenarios, covering *semantic context*, *visualization* and *interaction techniques*, as well as the *datasets*.

It is worth noting that most of the datasets used in our surveyed papers are not public standard datasets, but rather collected through the internet platforms or recorded by the authors. Datasets in the field of video visualization are usually designed for specific application processes, hence the data used not only includes the raw videos but needs to extract other modalities and forms of information extracted from these videos according to different task requirements to support visual analysis and exploration. Therefore, we summarize the datasets based on specific application scenarios.

### A. Surveillance Video

In various domains such as transportation [70], [96], [131], [146], residential facilities [66], [89], [107], and public safety [54], [88], [132], there is a growing demand for high-quality on-site videos and intelligent security alerts. Video surveillance techniques have been widely employed in societal production and daily life, particularly in the areas of security monitoring [45], [88], early warning [54], and emergency linkage [132]. However, the lengthy and monotonous nature of surveillance video poses great challenges for surveillance video review. The primary objective of surveillance video analysis is to perform *video summarization* and *identify video anomalies*. Existing approaches model and analyze the state sequences of active subjects (individuals and vehicles) in surveillance videos, achieving advanced semantic mining and comprehension of video content. In this section, we categorize these visual analysis approaches based on different surveillance environments: **outdoor-based** [53], [54], [85], [86], [132] and **indoor-based** [50], [88], [89]. We provided

| Application Scenario Description | Semantic Context | Visualization Technique | Interaction Technique | Dataset | Evaluation |
|---|---|---|---|---|---|
| **Surveillance video** is mainly divided into indoor and outdoor. *Outdoor surveillance video* has a wide detection range, low video quality, and lack of target detail information, so the track movement is the main information conveyed. *Indoor surveillance video* has a small detection range, high video quality, and supports various sensors and audio recording tools for monitoring. | • Object trajectory [53], [66], [83], [86], [88], [96], [131]<br>• Emotional information [3]<br>• Head movements [117]<br>• Audio information [2], [26], [115]<br>• Sensor information [54], [96], [117] | • chart-based [3], [26], [45], [53], [50]<br>• Graph-based [45]<br>• Timeline-based [2], [3], [26], [45], [53], [54], [50], [86], [96]<br>• 3D-based [53] | • Scaling [3], [26], [45], [53], [50]<br>• Filter [2], [3], [26], [45], [53], [54], [58], [50], [86], [96]<br>• Emphasis [45]<br>• Sketch [50] | • Public place videos [53], [83], [86]<br>• Shopping mall datasets [88]<br>• Tunnel videos [54]<br>• Campus videos [58]<br>• Classroom videos [3], [45]<br>• Mock online exam videos [50]<br>8min 30min | • **UWP** (50%)<br>• **VDAR** (70%)<br>• **UP** (40%)<br>• **UE** (80%)<br>• **AP** (40%) |
| **Sports videos** encompass a wide range of scenes, including dance, cycling, tennis, table tennis, snooker, baseball, soccer, and more. These videos serve as valuable sources of information, showcasing individual sports patterns and team tactics. | • Skeleton Information [27], [108]<br>• Target trajectory [60], [109], [106], [135], [136], [155]<br>• Target movement [4]–[6], [105], [126]<br>• Sensor information [150] | • Chart-based [5], [108], [148]<br>• Graph-based [148]<br>• Timeline-based [60], [109], [105], [108], [119], [128], [148]<br>• Projection-based [108]<br>• Glyph-based [4]–[6], [27], [85], [105], [128], [126], [135], [136]<br>• Sankey-based [60]<br>• 3D-based [27] | • Scaling [109]<br>• Filter [60], [85], [148], [153]<br>• Emphasis [105] | • Snooker videos [151]<br>• Baseball videos [153], [152]<br>• Tennis match videos [59], [118]<br>• Table tennis match videos [105]<br>• Volleyball videos [91]<br>• Soccer videos [85], [109], [135], [136], [148]<br>• Cycling videos [150]<br>• Dance videos [108], [149]<br>• Running videos [27]<br>30min 90min | • **UWP** (57.6%)<br>• **VDAR** (76.9%)<br>• **UP** (50%)<br>• **UE** (76.9%)<br>• **AP** (30.8%) |
| **Entertainment video** generally has distinct themes and rich narrative content, which is used to enrich people's spiritual life. The sources of entertainment videos are generally TV and Internet platforms. Such videos have high-definition video information and audio information. | • Emotional information [10]<br>• Text information [10], [11], [55], [71] | • Chart-based [55]<br>• Graph-based [10], [11], [51]<br>• Timeline-based [10], [11], [51], [55], [154]<br>• Projection-based [55]<br>• Glyph-based [10] | • Scaling [11], [51], [55]<br>• Filter [10], [11], [51], [55]<br>• Sketch [10] | • Movies [10], [11], [48], [71], [154]<br>• NHK's News [51]<br>• ARD news [55]<br>• TV program [71]<br>10min 120min | • **UWP** (50%)<br>• **VDAR** (87.5%)<br>• **UP** (50%)<br>• **UE** (87.5%)<br>• **AP** (12.5%) |
| **Educational videos** such as course videos, how-to videos and demonstration videos with HD video quality, containing audio and text information with lots of context and unstructured information. | • Target gestures [49]<br>• Target posture [7], [49]<br>• Emotional information [9], [101]<br>• Text information [74], [78] (e.g., Pop-ups, Subtitles, Captions, Script)<br>• Audio information [8], [9], [101], [122] | • Chart-based [8]<br>• Timeline-based [7], [9], [49], [74], [81], [101]<br>• Projection-based [9], [49], [101]<br>• Glyph-based [7], [9], [49]<br>• Sankey-based [7], [49] | • Scaling [7]–[9], [49]<br>• Filter [7]–[9], [49], [74], [78], [81], [47], [101]<br>• Emphasis [8], [9], [49] | • Speech videos [8], [9], [49], [101]<br>• Course videos [74], [81]<br>• Makeup videos [47]<br>• NPTEL dataset [79]<br>10min 90min | • **UWP** (15.4%)<br>• **VDAR** (84.6%)<br>• **UP** (53.8%)<br>• **UE** (100%)<br>• **AP** (15.4%) |

Fig. 17. Summary of the state-of-the-art papers from the perspective of application scenarios, covering *semantic context*, *visualization*, *interaction techniques*, *dataset*, and *evaluation*. The glyph in the dimension of the *dataset* represents the average time duration of the videos. Evaluation: understanding environments and work practices (UWP), visual data analysis and reasoning (VDAR), user performance (UP), user experience (UE), and algorithm performance (AP).

detailed descriptions of these two categories of research in the following subsections.

**Outdoor-based surveillance** involves complex scenes and objects, such as background disturbances, dynamically changing environments and the presence of non-target objects. The datasets involved in this scenario include public place videos [53], [85], [86], tunnel videos [54], and campus videos [58]. The duration of these videos is approximately 30 minutes. Current works major involves two aspects: *summarization of the object activity* and *anomalous monitoring*.

Certain studies focus on *summarizing the activity* of pedestrians in surveillance videos [53], [85], [96]. Due to the definition of outdoor-based surveillance video is low, and the monitored object occupying a small proportion of the video screen, it is difficult to capture microscopic human activities such as facial expressions, presentation content, and posture. Therefore, the trajectory information of moving objects has become the most critical feature in exploring the activities. Notably, spatial location information and temporal information are the main attributes of trajectory data.

In the stage of trajectory data extraction, the visual techniques involve moving object detection and trajectory matching. Most works [53], [85], [86] employ optical flow, foreground and background segmentation, and frame-difference methods to achieve moving object detection. Based on the detected objects, they locate the context position of the same target based on the proximity of object positions in adjacent frames to achieve trajectory extraction. This trajectory extraction method requires less computation. However, it is not suitable for scenes with high crowd density.

In terms of visual design, existing works represent trajectories with points [53], lines [86], and strips [85]. Nie et al. [86]

and Hoeferlin et al. [85] summarize the motion state of the monitored objects based on the trajectory data, to realize the fast browsing and retrieval of the video. Specifically, Nie et al. [86] propose a compact video synopsis technique based on optimized spatiotemporal trajectory. This technique achieves trajectory fusion without overlap and collision which improves the utilization of time and space. Hoeferlin et al. [85] map clustered trajectory data to a 2D visual space. They draw video scenes with cartoon illustrations and use strips with arrows to represent the trajectories. This form of visual mapping improves the speed of data perception and enhances the user's memory of video content. In order to enable users to interactively filter and explore trajectory patterns, Meghdadi et al. [53] map the trajectories into visual a 3D cube (Fig. 16(A)). The visual prototype they proposed allows users to select regions of interest and filter events based on the spatiotemporal characteristics of motion.

Furthermore, some studies are dedicated to *anomalous monitoring* of surveillance videos [54], [58], employing automated anomaly event detection algorithms for video surveillance. Fan et al. [58] (Fig. 16(B)) utilize IBM-S3 for temporal anomaly detection and employ a matrix arrangement video visualization approach that integrates information from different camera sensors and geographical locations to enhance detection accuracy and response speed. Piringer et al. [54] utilizes sequence analysis techniques to predict potential anomalies (such as fires) in tunnels, and employs tunnel imagery visualization with abstract graphics to visualize detected events. Users can access any temporal and spatial points to validate predicted anomalies in real-time or historical videos.

**Indoor-based Surveillance** has a relatively narrow monitoring scope and can be assisted with various sensors and audio recording tools (such as recorders, muscle sensors, and motion sensors.) to achieve more detailed monitoring. Effectively utilizing such data to achieve advanced semantic content understanding in videos is a promising research topic. A critical task is the transformation, alignment, and coordination of multimodal data. Another critical task is the transformation, alignment, and coordination of multimodal data. The datasets involved in this scenario include shopping mall datasets [88], classroom videos [3], [45] and mock online exam videos [50]. The duration of these videos scales from 10 minutes to 30 minutes. Compared with outdoor surveillance video, indoor surveillance video has a higher definition and can capture more subtle human activity information. Therefore, the subtle state information of the monitored object, such as facial emotion, posture, and motion state, becomes the main features in exploring the activities [3], [88], [89].

Most of the existing visual analysis techniques use timeline-based visualization to represent the state information of monitored objects, which aims to convey the evolution patterns over time. Zeng et al. [3] model facial emotion to mine and analyze student status in the classroom. They propose an interactive visual analysis tool EmotionCues, which uses multi-view linkage to present the evolution pattern of the monitored student's emotion. Likewise, Li et al. [50] consider head movement as an essential indicator for assessing the cheating behavior of candidates during online exams.

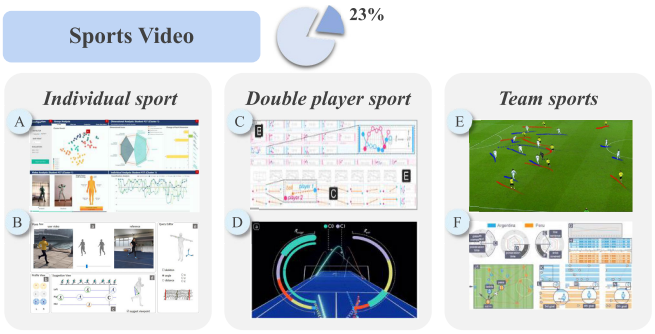



Fig. 18. (A) DanceVis [108] is proposed to assist teachers in evaluating the dance movements of students. (B) Liu et al. [106] proposed a viewpoint-invariant visualization approach that facilitates interactive and customized analysis of motion postures within videos. (C) Polk et al. [104] designed a visualization system named CourtTime to analyze tennis videos. (D) ShuttleSpace [147] visualized the shuttle trajectories in virtual reality; (E) Stein et al. [136] integrated multiple visualization charts into soccer videos for further analysis. (F) Wu et al. [148] proposed a visual analysis system, named ForVizor, to help users explore the team movements of soccer matches.

They also use band diagrams to plot the candidates' head movements at different angles. To associate multimodal data, Botchen et al. [88] (Fig. 16(D)) combine geometric information, semantic information, and statistical information to realize the mining of individual behavior patterns and the correlation of inter-individual behavior patterns. There are also some works devoted to analyzing overhead perspective videos. For example, Romero et al. [89] focus on analyzing surveillance video from the overhead view (Fig. 16(C)). They create a 3D activity cube analysis tool so that users can observe activity changes in a monitored space.

### B. Sports Video

Sports videos have been visualized and analyzed by various visual analysis methods to help people mine individual movement patterns, confrontation strategies, and team tactics. Current sports video visualizations mainly focus on specific sports scenarios, such as running [27], dance [108], [149], cycling [150], tennis [57], [104], table tennis [105], snooker [151], baseball [152], [153], rugby [60], and soccer [109]. The datasets for these videos are typically sourced from public sports matches [105], [109] or recorded by the authors [84], [148]. The duration of these videos scales from 30 minutes to 90 minutes. The view scope and clarity of these videos vary with different sports scenarios and the number of athletes involved, leading to differing complexities in visual analysis. Therefore, we further divide these sports into three categories: **individual sports**, **double-player sports**, and **team sports** based on the number of athletes.

**Individual sports** involve only one athlete, such as dance [108], [149], bicycling [150], and running [27]. The visual analysis techniques based on individual sports focus on the kinematic performance and technical intricacies of individual athletes. These analytical approaches aim to track the articulations and bodily segments of athletes, capturing and analyzing the precision of their postures, the fluidity of their movements, and the congruity of their technical

elements. They enable the visualization and analysis of action curves [149], motion trajectories [150], and other data [108].

By employing visualization techniques, a comprehensive understanding of athletes' technical strengths and areas for improvement can be obtained, thereby providing them with more precise training guidance. For example, DanceVis [108] is proposed to assist teachers in evaluating the dance movements of students. It extracts four physical and four behavioral attributes from dance video. Then, it calculates similarities between standard and student pose skeletons as the recommended score of their dance. In visualization, they integrate multiple charts such as scatter plots, radar charts, bar charts, and line charts to present the reference indicator in multi-aspect for teachers. In order to enable amateur runners to discern the fundamental disparities between their running postures and those of professional runners, Liu et al. [27] (Fig. 18(B)) propose a viewpoint-invariant approach that facilitates interactive and customized analysis of motion postures within videos.

Moreover, there are also some techniques devoted to enhance the kinematic patterns of crucial body segments by incorporating graphical embeddings into individual sports. Kaplan et al. [150] propose a visual design that embeds visual representations into cycling training videos to enhance the pattern of cyclists' pedaling. In this work, the authors apply circular form charts (e.g., circle arrows and triangles) to display the pedaling data to help users figure out and understand cyclists' movement patterns in cycling training. To further enhance the emotional expression in the such videos, Payne et al. [149] (Fig. 18(A)) propose danceOn, which allows users to design cartoon elements based on different movement patterns and embed them into dance videos to enhance the emotional expression.

**Double-player sports** involve two players such as badminton [137], [147], tennis [59], [104], [118], table tennis [57], [105], [128], and snooker [151]. The visual analysis techniques based on double-player sports focuses on the interaction or adversarial relationship between two participating athletes. The primary objective of these techniques are to delve into the strategies of athletes' movements. In terms of visualization, the main emphasis contents are placed on the visual representation of spatial relationships [137], [147], confrontational movements [59], [104], and modes of mobility between the objects [105]. Through the implementation of visual analysis techniques, one can observe and study the tactical choices, execution, and resultant effects within the domain of double-player sports. The invaluable value of such endeavors extends to coaches and athletes alike, as it facilitates collaborative improvements, fortifies opponent analysis, and enables the formulation of more efficacious strategies. However, there are two major challenges: (1) how to effectively present the overview of match details and (2) how to efficiently annotate large and dense motion events in sports videos.

To analyze and summarize match detail in double player sports effectively, Ye et al. [147] proposed an immersive visual analysis tool named ShuttleSpace to help analysts explore and analyze the movement of ball trajectory in badminton games (Fig. 18(D)). Similarly, Polk et al. [104] (Fig. 18(C)) designed a visualization system named CourtTime to analyze tennis videos. In this work, at first, they collect the location data from the match videos. Then, 1D space-time charts and 2D movement charts are employed to convey the insights of the tennis match. In addition, researchers work on reducing human interaction and effort in annotating double player sports videos [126], [128]. For example, VisCommentator [126] automatically extracts the objects and events in table tennis videos and allows users to interactively annotate these items. It allows analysts to brush the timeline and select recommended visual elements to generate table tennis augmenting videos.

**Team sports** are usually played with two teams involving many athletes such as soccer [106], [109], American football [60], [135], and baseball [119], [152], [153]. For team sports videos, research focuses on retrieving and summarizing the team movement pattern and analyzing team tactics. In the context of team sports, the visual analysis techniques focus on analyzing the collective behavior [60], [136], team tactics [106], [109], and overall performance [135] in team sports. When it comes to visual analysis, the focus lies on presenting the positional relationships [60], [106], [109] and movement trajectories [135], [136] among multiple individuals. Through visual analysis, one can observe and analyze team tactics, defensive strategies, and offensive organization within team sports. This holds crucial significance for training, tactical planning, and strategic formulation in team sports, ultimately enhancing overall collaboration and performance levels. However, how to precisely detect, present, and summarize team sports strategies becomes a challenge due to the rapidly changing motion trajectories and the difficulty in tracking moving targets.

To effectively summarize the team tactics, Wu et al. [148] (Fig. 18(F)) propose a visual analysis system, named ForVizor, to help users explore the team movements of soccer matches. They utilize the proposed automatic algorithms to detect the team movements automatically. Based on this detected information, they use multiple linked views such as matrix, narrative timeline, and pitch to help users dig out the formations and tactics transformation of teams. To visualize the team movement patterns, Stein et al. [136] (Fig. 18(E)) incorporated movement data into the raw video to monitor sports patterns in real time. In visualization, the technniques embed many visualization charts such as heat maps, doughnut charts, and timeline movement paths into videos to present the dominant regions, pass distances, players' movements, and players' reactions in soccer matches.

### C. Entertainment Video

Entertainment video generally has distinct themes and rich narrative content. The datasets of entertainment videos are generally movies [10], [48], [154], TV programs [71], news [51], [55], etc. obtained from online platforms. The duration of these videos scales from 10 minutes to 120 minutes. Such videos have high-definition video information and audio information. Therefore, this type of video data is usually high-quality and multimodal. Regarding the analysis of entertainment videos, existing research primarily concentrates on *summarizing* and *comprehending* medium-sized entertainment

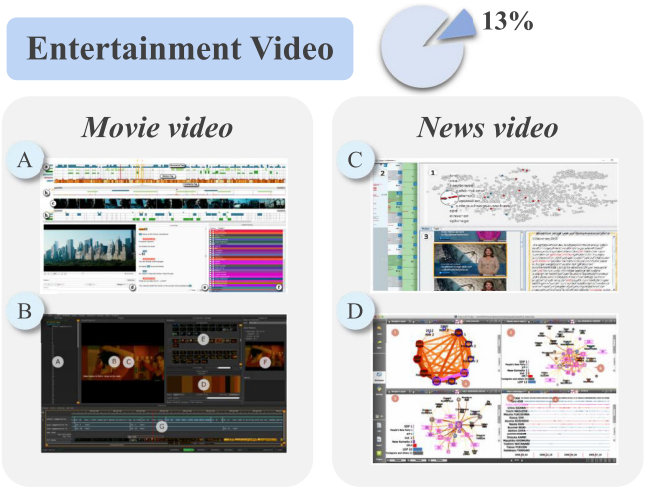



Fig. 19. (A) Kurzhals et al. [48] performed semantic segmentation of the movie's content to conduct a multi-level semantic inference analysis of the video content. (B) VIAN [154] summarized the video content from the aspect of color. (C) Markus et al. [55] summarized the news topics by extracting the text in news videos. (D) Renoust et al. [51] explored patterns of association between politicians by face tracking.

videos with compact content. This analysis assists users in enhancing their video-watching efficiency or providing tailored recommendations for captivating segments. In this section, we mainly review the visual analysis work based on **movie videos** [10], [48], [71], [154] and **news videos** [51], [55],which represent the two most prevalent types of entertaining videos.

**Movie video** is an artistic creation that revolves around the elements of storytelling, performances, and visual effects. Specifically, these contents are reflected in the style, theme, emotion, character relationships in the movie, as well as the director's creative techniques and intentions. However, the perception of art forms such as movies tends to be subjective, which poses challenges in precisely defining analysis tasks and establishing quantifiable evaluation criteria for movie videos. Furthermore, movie videos encompass various multimodal information, including images, audio, and scripts, thereby amplifying the complexity of data and enhancing the difficulties encountered in visual analysis. To address these issues, the existing techniques have been carried out such as scene analysis [48], [71], emotion recognition [10], shot composition analysis [11], and visual effects analysis [154], enabling a comprehensive understanding and assessment of movies. Existing visualization techniques are introduced to explore movie videos from two major aspects: *movie content* [10], [48], [71] and *movie styles* [11], [154].

*Movie content* can be summarized into four key elements: *when*, *where*, *who*, and *what*. Existing research often analyzes the movie content from one or several specific elements. EmotionMap [10] proposes EmotionDisc to model emotions based on the emotion detection techniques of face, text, and audio. Additionally, it designs a compact map-style visualization that integrates the information consisting of time sequence, character emotions, events, and the correlations between items to summarize the semantic structure of the movie. In order to analyze the movie from a more macro perspective, some works analyze the movie content in scenes [48], [71]. Kurzhals et al. [48](Fig. 19(A)) perform semantic segmentation of the movie content based on integrated multimodal information (including script, images, scripts, and subtitles) to conduct multi-level semantic analysis of the video content. In visualization, their work provides a scalable timeline that allows users to freely filter and refine the presented video granularity.

*Movie style* includes shooting techniques such as the application of color and scene transitions. To analyze movie style from the perspective of color, VIAN [154] (Fig. 19(B)) combines background segmentation techniques with human perception of color to assess the shooting style and aesthetic quality of the movie. VideoForest [11] generates session-based video summaries by leveraging bullet screen data and video frame information, and enable the exploration of shooting techniques and filter styles. In the visual design of summaries, it introduces a forest-themed visualization approach to metaphorically showcase movie scenes and keyframes.

**News videos** has characteristics of concise narratives and clear content structures. The focal points of news video analysis often encompass aspects such as accuracy, objectivity, reporting style, and news value, aimed at evaluating the quality and credibility of news reporting. Analyzing news videos allows for the exploration of content, themes, perspectives, and linguistic styles conveyed in news reports, as well as assessing the impact and effectiveness of news dissemination. This analysis may involve techniques such as speech recognition, sentiment analysis, keyword extraction, and event detection. Compared to movies, news videos employ simpler shooting techniques, with camera angles switching among a few fixed positions. Based on the specific news scenario, news videos can be categorized as *studio news* [55] and *on-site news* [51].

For *studio news*, the video scene is relatively stable, and key information may be reflected in text and audio information. Markus et al. [55] (Fig. 19(C)) employ optical character recognition technology to extract subtitle and title information from video images. They then utilize topic extraction techniques to semantically cluster news in a collection of videos. Moreover, they project news data using topic-based vectors to compare the similarities and dissimilarities between different topics. Additionally, they utilize a multi-level timeline to summarize the evolution patterns of news content at varying scales, facilitating user filtering and querying. For *on-site news* which is recorded in real scenes, the characters in the video may be the focus of the news. Renoust et al. [51] (Fig. 19(D)) propose a political analysis visualization system for a large-scale news video archive based on facial tracking. This system utilizes facial detection and tracking techniques to construct a political network, aiding users in gaining a deeper understanding of political interactions and media phenomena through four levels of abstraction: time segments, networks, timeline, and facial tracking within the videos. Moreover, in terms of visualization, this research examines the patterns of appearances and relationships among politicians in a graph-based visualization, enabling users to selectively explore individual network connections.

### *D. Education Video*

Education videos contain a wealth of contextualized and unstructured information. The datasets of education videos

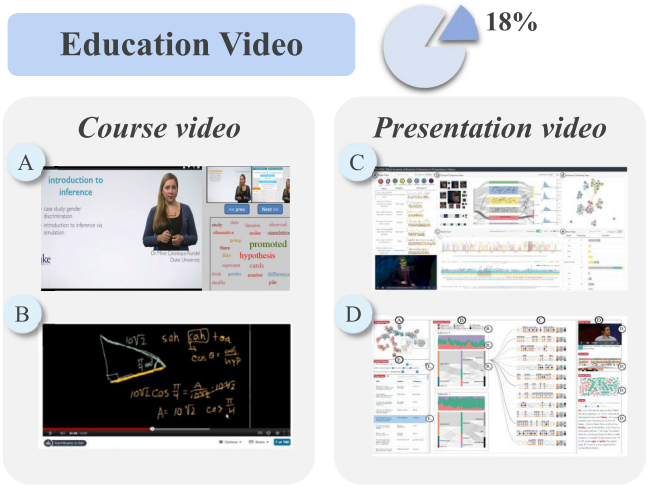



Fig. 20. (A) Yadav et al. [74] designed a dynamic timeline to support nonlinear navigation. (B) Monserrat et al. [81] extracted the concepts in blackboard-style instructional video. (C) EmoCo [9] analyzed the consistency and evolution of emotions in presentation video. (D) Wu et al. [49] analyzed the concordance between verbal and non-verbal in presentation video.

are generally presentation videos [8], [9], [101] and course videos [74], [79], [81] obtained from internet platforms such as YouTube and MOOC. The duration of these videos scales from 10 minutes to 90 minutes. The main purpose of analyzing these education videos is primarily in swifting knowledge acquisition and enhancing the quality of online teaching. These videos typically often have high-definition visuals and audio components, necessitating a comprehensive analysis encompassing textual, auditory, and visual content. In this section, we further categorize these educational videos into **course videos** [73], [102] and **presentation videos** [9], [101].

**Course video analysis** techniques focus primarily on the *instructional content* [74], [79], [81],and *learning behavior analysis* [102], [155], [156], [157]. The *instructional content* in course videos often exhibits a monotonous and repetitive nature, which may lead to reduced student concentration and diminished patience. Students are required to repetitively seek out and view specific video segments in order to enhance their comprehension of the knowledge points. To promote efficient browsing and learning, some works [47], [79], [81] extract keywords and keyframes from videos for navigation. Yadav et al. [74] (Fig. 20(A)) propose a multi-dimensional nonlinear video navigation tool that utilizes blackboard information and audio data. Regarding visualization, it employs the dynamic time-aware word clouds and key point timelines to facilitate users in precisely locating specific points of interest or key segments within the video. In order to obtain semantic information to navigate, Monserrat et al. [81] (Fig. 20(B)) extract the concept of geometric shapes, and formulas from blackboard-style education videos and create a summary image of video concepts as a navigation interface, enabling users to directly navigate to corresponding video frames of specific concepts. Truong et al. [47] utilize principles from cognitive psychology on how humans perceive, remember, and communicate event structures to automatically extract a two-level hierarchical overview of makeup instructional tutorials. They consider facial elements as high-level events while treating actions as low-level micro-objects.

Some studies are dedicated to conducting *learning behavior analysis* of students while they watch course videos. These studies [102], [155], [156] devote to gain insights into students' behavioral patterns and learning modalities, with the aim of evaluating pedagogical effectiveness and uncovering learning difficulties of the course video, thus facilitating a comprehensive understanding of students' learning behaviors and individual preferences. The VisMOOC [156] analyzed video clickstream data on MOOC platforms. Based on content-based views, it displays the time variation of the total number of each type of click action on the video timeline to help MOOC instructors analyze user learning behavior. However, discovering the utilization pattern of a massive number of videos from learning log data remains a significant challenge.

**Presentation videos** demand learners to engage in detailed analysis and exploration, aiming to uncover hidden insights, such as humor skills or the establishment of emotional resonance within the presentation. Nonetheless, the dig of presentation often suffers from a lack of precise definitions and uniform evaluation criteria, which frequently confuse novices delving into presentation content. In order to solve these problems, some works use visual analysis techniques to help users understand the complexity of presentation videos. The main analysis contents include language style [8], body posture [9], [49] and personal emotions Tendency [101].

To learn the expression skills of humor, Dehumor [8] is designed to analyze the language style of the speaker from two aspects: (1) the interaction of vocal delivery and script information with inline annotations. (2) the contextual linking of concepts conveyed in the presentation with a context-linking graph. However, this work analyzes video only from audio data and ignores image information. Some works analyze image information of gestures and emotions in presentation expression [9], [49], [101]. EmoCo [9] is proposed to analyze the consistency and evolution of multimodal emotions in presentation videos (Fig. 20 (C)). They extract the face and detect the emotion of the speaker. In order to help users with multi-level analysis and exploration, this work allows users to analyze the multimodal sentiment (i.e., the sentiment of text, audio, and face) in presentation videos from three levels (video-level, sentence-level, and word-level) for comparative analysis. In visualization, this work designs a novel Sankey diagram to express the flow of emotions between different modalities. Furthermore, Wu et al. [49] (Fig. 20 (D)) analyze the concordance between verbal and non-verbal information in a collection of presentation videos. On the basis of this distilled information, they guide users to explore the video collection from three levels: video collections, video comparison in time series, and detail of a specific video. They utilize an innovative glyph to represent the postures in the presentation videos, along with employing scalable timeline navigation to facilitate users in conducting interactive exploration at varying levels of granularity.

## VI. Discussion and Outlook

In this section, we discuss video visualization and visual analysis from four aspects including scalability, uncertainty and evaluation, multimodal analysis, and real-time analysis.

### A. Scalability

With the rapid expansion of video data, video visualization, and visual analysis face significant challenges. Video data has the characteristics of substantial volume, redundancy, and complexity. Therefore, visual confusion often occurs during the visual mapping process. To address this issue, numerous existing research efforts employ frame sampling [158] or keyframe extraction [65] during the data processing stage. In the visual design, visual encoding techniques [101], [159] are widely employed to convey semantic information in the videos. Selecting appropriate graphic symbols to express video features can effectively minimize visual clutter and confusion. In the realm of visualization interaction, various approaches such as filtering, querying, and scaling are extensively employed to facilitate scalable analysis.

However, the current visual analysis techniques still have limitations in scalability. Their analytical capabilities limited to video content of up to two hours (as shown in Fig. 17). This phenomenon highlights the challenges in visual analysis of long video content. Additionally, the issue of balancing information presentation and visual space utilization remains insufficiently explored. Within a limited screen scope, determining the granularity and information density of video content that best facilitates human exploration remains an open question. These unresolved challenges also motivated us to undertake this review study. Future work could explore the development of more efficient methods for high visual throughput image representation, abstract representation of semantic information, and spatiotemporal integrated visual representation, aiming to compress the high-density pixel information in video data. Additionally, it might consider progressively expanding the presentation of information based on human intent to prevent overwhelming users at the outset.

### B. Uncertainty and Evaluation

Uncertainty in video analysis typically consists of two aspects: (1) Uncertainty in algorithms: Current efforts aim to construct a visual analysis system that integrates machine intelligence with human perceptions. However, machine learning algorithms are often perceived as a "black box" by developers of visualization tools and users [116], [117], making the deployment in practical applications prone to unexpected errors. Currently, there is a lack of in-depth research on enhancing users' trust in the results of video analysis models. To address this issue, interactive analysis and feedback should be introduced to incorporate human insights into the process of improving model outcomes.

(2) Uncertainty in data: Video data may contain noise, missing information, redundancy, or errors, which can introduce uncertainty in the analysis and interpretation of video content. In addition to efforts to repair and enhance uncertain data, visualization techniques can assist in addressing issues related to low-quality data. For example, conveying uncertainty information to users through visual representations such as uncertainty ranges, confidence charts, and fuzzy sets can facilitate accurate decision-making. Furthermore, visual analytics techniques can present the key factors and influences contributing to uncertainty, thereby promoting a deeper understanding of the analysis results.

The uncertainty creates challenges for accurately analyzing and understanding videos. Existing research methods [45], [108] mainly focus on pre-defined visual analysis of video information in specific domains, and transform uncertain video semantic information into specific analysis tasks. However, we still have the opportunity to explore personalized and customized exploration mechanisms to deal with more uncertainties. For example, the ability to analyze semantic uncertainty in videos can be further improved by employing methods such as gesture-based video information retrieval and context-based pattern mining.

Additionally, effective evaluation methods can help in understanding and quantifying these uncertainties, which contributes to the enhancement of the effectiveness and credibility of video visual analysis techniques. The comparative analysis in Fig. 2 reveals that little research regarding the performance evaluation of algorithms, reflecting the evaluation of video visualization techniques focuses on the effectiveness of the human-centered analysis process rather than the performance of quantitative metrics. Furthermore, since the algorithms used in video visualization techniques are often closely related to their application contexts, it becomes particularly challenging to establish a universal standard or metric to measure the performance of different algorithms.

### C. Multimodal Analysis

Multimodal data analysis has emerged as a prominent research topic in the field of video analysis [160], [161]. In existing work, semantic-level alignment and fusion of multimodal data constitute the primary methods employed for multimodal data processing. Such as integrating information from different modalities onto a common attribute dimension for alignment [9], or directly utilizing multimodal models for event or object recognition. Regarding visualization, prior research has presented simultaneous visual representations of diverse modalities, employing distinct visual symbols to depict the characteristics of each modality [7]. Futhermore, visualization forms such as Sankey diagrams and association graphs have been employed to reveal distinctions and consistencies among modalities.

Additionally, with the emergence of large language models such as ChatGPT [162], the field of computer vision and multimodal analysis has experienced significant advancements [163], [164]. These large models incorporate robust semantic understanding and expressive capabilities which brings opportunities for the combination of visualization techniques and automatic vision techniques. Particularly, these techniques stimulated research on previously challenging tasks [165]. For example, automated models extract high-level semantic information from videos, while visualization techniques can show the relationship between questions and

answers through intuitive visual representations to help users better understand the basis and reasoning of the answers.

### D. Real-Time Analysis

Current visual analytics [6], [106] efforts primarily focus on modeling historical video streams, transforming unstructured video data into structured formats to facilitate further pattern exploration. In contrast to offline analysis, analyzing online video streams aims to enable real-time processing and analysis of video streams, extracting valuable information for prompt decision-making [1]. However, this is a challenging task. On one hand, there is a high demand for real-time model performance. The diversity and complexity of video data further amplify the challenges associated with algorithm design and optimization. On the other hand, owing to the dynamic nature of online videos, upcoming video streams may contain content that is irrelevant to the current analysis task and filled with unknown variables, rendering predictions difficult. There is an urgent need to develop methods for more efficient processing and summarization of dynamic video stream data. Moreover, effectively presenting vast historical data, real-time updated data, and their interconnections within the field of visual analytics requires further research and in-depth exploration.

## VII. Conclusion

In this paper, we review the visualization and visual analytics for video data and provide a comprehensive overview. We first provide a design space based on the video visualization process. We then review and classify these papers from two dimensions: visual analysis tasks and applications. Specifically, visual analysis tasks are further divided into five types: video summarization, video content understanding, video anomaly detection, video editing, and video enhancement. The application scenarios are divided into the following four categories: surveillance, sports, entertainment, and education. In addition, our paper discusses the challenges and future research trends. This survey aims to provide insights to practitioners in this research direction and to help them better understand the role of visualization techniques in the process of exploring and analyzing video data.

## References


[1] A. D. Silva et al., "RipViz: Finding rip currents by learning pathline behavior," *IEEE Trans. Vis. Comput. Graphics*, vol. 1, no. 1, pp. 1–13, Nov. 2023.

[2] T. Tang, Y. Wu, Y. Wu, L. Yu, and Y. Li, "VideoModerator: A risk-aware framework for multimodal video moderation in e-commerce," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 846–856, Jan. 2022.

[3] H. Zeng et al., "EmotionCues: Emotion-oriented visual summarization of classroom videos," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 7, pp. 3168–3181, Jul. 2021.

[4] Z. Chen et al., "IBall: Augmenting basketball videos with gaze-moderated embedded visualizations," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2023, pp. 1–18.

[5] T. Lin, Z. Chen, Y. Yang, D. Chiappalupi, J. Beyer, and H. Pfister, "The quest for: Embedded visualization for augmenting basketball game viewing experiences," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 1, pp. 962–971, Jan. 2023.

[6] Z. Chen et al., "Sporthesia: Augmenting sports videos using natural language," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 1, pp. 918–928, Jan. 2023.

[7] H. Zeng, X. Wang, Y. Wang, A. Wu, T.-C. Pong, and H. Qu, "GestureLens: Visual analysis of gestures in presentation videos," *IEEE Trans. Vis. Comput. Graph.*, vol. 1, no. 1, pp. 1–14, Apr. 2022.

[8] X. Wang, Y. Ming, T. Wu, H. Zeng, Y. Wang, and H. Qu, "DeHumor: Visual analytics for decomposing humor," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 12, pp. 4609–4623, Dec. 2022.

[9] H. Zeng et al., "EmoCo: Visual analysis of emotion coherence in presentation videos," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 927–937, Jan. 2020.

[10] C.-X. Ma, J.-C. Song, Q. Zhu, K. Maher, Z.-Y. Huang, and H.-A. Wang, "EmotionMap: Visual analysis of video emotional content on a map," *J. Comput. Sci. Technol.*, vol. 35, no. 3, pp. 576–591, May 2020, doi: 10.1007/s11390-020-0271-2.

[11] Z. Sun, M. Sun, N. Cao, and X. Ma, "VideoForest: Interactive visual summarization of video streams based on danmu data," in *Proc. SIGGRAPH ASIA Symp. Visualizat.*, Nov. 2016, pp. 1–8.

[12] Y. Wang, M. Liu, J. Wu, and L. Nie, "Multi-granularity interaction and integration network for video question answering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 1–16, Nov. 2023.

[13] W. Luo et al., "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1070–1084, Mar. 2021.

[14] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2293–2312, May 2022.

[15] C. Chen, H. Wang, Y. Fang, and C. Peng, "A novel long-term iterative mining scheme for video salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7662–7676, Nov. 2022.

[16] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 994–1000.

[17] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[18] J. Cao, J. Pang, X. Weng, R. Khirodkar, and K. Kitani, "Observation-centric SORT: Rethinking SORT for robust multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Jun. 2023, pp. 9686–9696.

[19] Z. Zhou, X. Zhou, Z. Chen, P. Guo, Q.-Y. Liu, and W. Zhang, "Memory network with pixel-level spatio-temporal learning for visual object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 1–17, Dec. 2023.

[20] A. Kirillov, Y. Wu, K. He, and R. Girshick, "PointRend: Image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9799–9808.

[21] Y.-P. Zhao, X. Dai, Z. Wang, and X. Li, "Subspace clustering via adaptive non-negative representation learning and its application to image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 1–13, Sep. 2023.

[22] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2021.

[23] M. Yao, D. He, X. Li, F. Li, and Z. Xiong, "Towards interactive self-supervised denoising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 1–24, Sep. 2023.

[24] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[25] H. Wu et al., "DisCoVQA: Temporal distortion-content transformers for video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 1–15, Sep. 2023.

[26] K. K. Wong, X. Wang, Y. Wang, J. He, R. Zhang, and H. Qu, "Anchorage: Visual analysis of satisfaction in customer service videos via anchor events," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 8, pp. 1–13, Oct. 2023.

[27] J. Liu et al., "PoseCoach: A customizable analysis and visualization system for video-based running coaching," *IEEE Trans. Vis. Comput. Graphics*, vol. 1, no. 1, pp. 1–14, Oct. 2022.

[28] S. Afzal et al., "Visualization and visual analytics approaches for image and video datasets: A survey," *ACM Trans. Interact. Intell. Syst.*, vol. 13, no. 1, pp. 1–41, Mar. 2023.

[29] G.-D. Sun, Y.-C. Wu, R.-H. Liang, and S.-X. Liu, "A survey of visual analytics techniques and applications: State-of-the-art research and future challenges," *J. Comput. Sci. Technol.*, vol. 28, no. 5, pp. 852–867, Sep. 2013.

[30] G. Sun et al., "Application of mathematical optimization in data visualization and visual analytics: A survey," *IEEE Trans. Big Data*, vol. 9, no. 4, pp. 1–20, Sep. 2023.

[31] N. A. Chinchor, J. J. Thomas, P. C. Wong, M. G. Christel, and W. Ribarsky, "Multimedia analysis+ visual analytics= multimedia analytics," *IEEE Comput. Graph. Appl.*, vol. 30, no. 5, pp. 52–60, Sep. 2010.

[32] R. Borgo et al., "A survey on video-based graphics and video visualization," *Eurographics*, vol. 1, pp. 1–23, Jul. 2011.

[33] B. Höferlin, M. Höferlin, G. Heidemann, and D. Weiskopf, "Scalable video visual analytics," *Inf. Visualizat.*, vol. 14, no. 1, pp. 10–26, Jan. 2015.

[34] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proc. IEEE*, vol. 109, no. 11, pp. 1838–1863, Nov. 2021.

[35] P. Meena, H. Kumar, and S. Kumar Yadav, "A review on video summarization techniques," *Eng. Appl. Artif. Intell.*, vol. 118, Feb. 2023, Art. no. 105667.

[36] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1993–2008, Nov. 2013.

[37] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[38] M. T. Fischer, D. A. Keim, and M. Stein, "Video-based analysis of soccer matches," in *Proc. 2nd Int. Workshop Multimedia Content Anal. Sports*, Oct. 2019, pp. 1–9.

[39] J. Wang, T. Gui, M. Cheng, X. Wu, R. Ruan, and M. Du, "A survey on emotional visualization and visual analysis," *J. Visualizat.*, vol. 26, no. 1, pp. 177–198, Feb. 2023.

[40] H.-C. Shih, "A survey of content-aware video analysis for sports," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1212–1231, May 2018.

[41] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 367–386, Mar. 2015.

[42] K. Schoeffmann, M. A. Hudelist, and J. Huber, "Video interaction tools: A survey of recent work," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 1–34, Sep. 2015.

[43] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1520–1536, Sep. 2012.

[44] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A systematic review on the practice of evaluating visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2818–2827, Dec. 2013.

[45] G. Sun, T. Li, and R. Liang, "SurVizor: Visualizing and understanding the key content of surveillance videos," *J. Visualizat.*, vol. 25, no. 3, pp. 635–651, Jun. 2022.

[46] B. Huber, H. V. Shin, B. Russell, O. Wang, and G. J. Mysore, "B-Script: Transcript-based B-roll video editing with recommendations," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–11.

[47] A. Truong, P. Chi, D. Salesin, I. Essa, and M. Agrawala, "Automatic generation of two-level hierarchical tutorials from instructional makeup videos," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–16.

[48] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf, "Visual movie analytics," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2149–2160, Nov. 2016.

[49] A. Wu and H. Qu, "Multimodal analysis of video collections: Visual exploration of presentation techniques in TED talks," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 7, pp. 2429–2442, Jul. 2020.

[50] H. Li, M. Xu, Y. Wang, H. Wei, and H. Qu, "A visual analytics approach to facilitate the proctoring of online exams," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–17.

[51] B. Renoust, D.-D. Le, and S. Satoh, "Visual analytics of political networks from face-tracking of news video," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2184–2195, Nov. 2016.

[52] S. Jang, N. Elmqvist, and K. Ramani, "MotionFlow: Visual abstraction and aggregation of sequential patterns in human motion tracking data," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 21–30, Jan. 2016.

[53] A. H. Meghdadi and P. Irani, "Interactive exploration of surveillance video through action shot summarization and trajectory visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2119–2128, Dec. 2013.

[54] H. Piringer, M. Buchetics, and R. Benedik, "AlVis: Situation awareness in the surveillance of road tunnels," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2012, pp. 153–162.

[55] M. John, K. Kurzhals, and T. Ertl, "Visual exploration of topics in multimedia news corpora," in *Proc. 23rd Int. Conf. Inf. Vis. (IV)*, Jul. 2019, pp. 241–248.

[56] H. Liang, R. Liang, and G. Sun, "Looking into saliency model via space-time visualization," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2271–2281, Nov. 2016.

[57] Y. Wu et al., "ITTVis: Interactive visualization of table tennis data," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 709–718, Jan. 2018.

[58] C.-T. Fan, Y.-K. Wang, and C.-R. Huang, "Heterogeneous information fusion and visualization for a large-scale intelligent video surveillance system," *IEEE Trans. Syst. Man, Cybern., Syst.*, vol. 47, no. 4, pp. 593–604, Apr. 2017.

[59] T. Polk, J. Yang, Y. Hu, and Y. Zhao, "TenniVis: Visualization for tennis match analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 2339–2348, Dec. 2014.

[60] P. A. Legg et al., "Transformation of an uncertain video search pipeline to a sketch-based visual analytics loop," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2109–2118, Dec. 2013.

[61] C.-X. Ma, Y.-J. Liu, G. Zhao, and H.-A. Wang, "Visualizing and analyzing video content with interactive scalable maps," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2171–2183, Nov. 2016.

[62] C.-X. Ma, Y.-J. Liu, H.-A. Wang, D.-X. Teng, and G.-Z. Dai, "Sketch-based annotation and visualization in video authoring," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1153–1165, Aug. 2012.

[63] C.-X. Ma, Y. Guo, and H.-A. Wang, "VideoMap: An interactive and scalable visualization for exploring video content," *Comput. Vis. Media*, vol. 2, no. 3, pp. 291–304, Sep. 2016.

[64] M. Chang, M. Huh, and J. Kim, "RubySlippers: Supporting content-based voice navigation for how-to videos," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–14.

[65] K. Subramanian, J. Maas, J. Borchers, and J. Hollan, "From detectables to inspectables: Understanding qualitative analysis of audiovisual data," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–10.

[66] Y. Wang et al., "Effects of video placement and spatial context presentation on path reconstruction tasks with contextualized videos," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1755–1762, Nov. 2008.

[67] K. Sunkavalli, N. Joshi, S. B. Kang, M. F. Cohen, and H. Pfister, "Video snapshots: Creating high-quality images from video clips," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 11, pp. 1868–1879, Nov. 2012.

[68] Y. Jung, D. Cho, S. Woo, and I. S. Kweon, "Global-and-local relative position embedding for unsupervised video summarization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 167–183.

[69] W. Zhu, Y. Han, J. Lu, and J. Zhou, "Relational reasoning over spatial-temporal graphs for video summarization," *IEEE Trans. Image Process.*, vol. 31, pp. 3017–3031, 2022.

[70] M. Höferlin, K. Kurzhals, B. Höferlin, G. Heidemann, and D. Weiskopf, "Evaluation of fast-forward video visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2095–2103, Dec. 2012.

[71] T. Chen, A. Lu, and S.-M. Hu, "Visual storylines: Semantic visualization of movie sequence," *Comput. Graph.*, vol. 36, no. 4, pp. 241–249, Jun. 2012.

[72] S. Samrose et al., "MeetingCoach: An intelligent dashboard for supporting effective & inclusive meetings," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–13.

[73] J. Kim, P. T. Nguyen, S. Weir, P. J. Guo, R. C. Miller, and K. Z. Gajos, "Crowdsourcing step-by-step information extraction to enhance existing how-to videos," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2014, pp. 4017–4026.

[74] K. Yadav et al., "Content-driven multi-modal techniques for non-linear video navigation," in *Proc. 20th Int. Conf. Intell. User Interface*, Mar. 2015, pp. 333–344.

[75] A. Al-Hajri, G. Miller, M. Fong, and S. S. Fels, "Visualization of personal history for video navigation," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2014, pp. 1187–1196.

[76] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala, "Video digests: A browsable, skimmable format for informational lecture videos," in *Proc. 27th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2014, pp. 573–582.

[77] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 510–522, Oct. 2010.

[78] J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash, and L. A. Rowe, "TalkMiner: A lecture webcast search engine," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 241–250.

[79] A. Biswas, A. Gandhi, and O. Deshmukh, "MMToC: A multimodal method for table of content creation in educational videos," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 621–630.

[80] C. A. Fraser, T. J. Ngoon, M. Dontcheva, and S. Klemmer, "RePlay: Contextually presenting learning videos across software applications," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2019, pp. 1–13.

[81] T.-J.-K. Monserrat, S. Zhao, K. Mcgee, and A. V. Pandey, "NoteVideo: Facilitating navigation of blackboard-style lecture videos," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2013, pp. 1139–1148.

[82] J. Kim, "Toolscape: Enhancing the learning experience of how-to videos," in *Proc. CHI Extended Abstr. Hum. Factors Comput. Syst.*, Apr. 2013, pp. 2707–2712.

[83] M. E. Swift, W. Ayers, S. Pallanck, and S. Wehrwein, "Visualizing the passage of time with video temporal pyramids," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 1, pp. 171–181, Jan. 2023.

[84] R. Hamid, R. Kumar, J. Hodgins, and I. Essa, "A visualization framework for team sports captured using multiple static cameras," *Comput. Vis. Image Understand.*, vol. 118, pp. 171–183, Jan. 2014.

[85] M. Höferlin, B. Höferlin, G. Heidemann, and D. Weiskopf, "Interactive schematic summaries for faceted exploration of surveillance video," *IEEE Trans. Multimedia*, vol. 15, no. 4, pp. 908–920, Jun. 2013.

[86] Y. Nie, C. Xiao, H. Sun, and P. Li, "Compact video synopsis via global spatiotemporal optimization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 10, pp. 1664–1676, Oct. 2013.

[87] W.-S. Lai, Y. Huang, N. Joshi, C. Buehler, M.-H. Yang, and S. B. Kang, "Semantic-driven generation of hyperlapse from 360 degree video," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 9, pp. 2610–2621, Sep. 2018.

[88] R. P. Botchen et al., "Action-based multifield video visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 4, pp. 885–899, Jul. 2008.

[89] M. Romero, J. Summet, J. Stasko, and G. Abowd, "Viz-A-vis: Toward visualizing video through computer vision," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 6, pp. 1261–1268, Nov. 2008.

[90] J. Cao, C.-W. Ngo, Y.-D. Zhang, and J.-T. Li, "Tracking web video topics: Discovery, visualization, and monitoring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1835–1846, Dec. 2011.

[91] H.-T. Chen, W.-J. Tsai, S.-Y. Lee, and J.-Y. Yu, "Ball tracking and 3D trajectory approximation with applications to tactics analysis from single-camera volleyball sequences," *Multimedia Tools Appl.*, vol. 60, no. 3, pp. 641–667, Oct. 2012.

[92] G. Y. Chan, L. G. Nonato, A. Chu, P. Raghavan, V. Aluru, and C. T. Silva, "Motion browser: Visualizing and understanding complex upper limb movement under obstetrical brachial plexus injuries," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 981–990, Jan. 2020.

[93] B. Duffy et al., "Glyph-based video visualization for semen analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 8, pp. 980–993, Aug. 2015.

[94] K. Kurzhals and D. Weiskopf, "Space-time visual analytics of eye-tracking data for dynamic stimuli," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2129–2138, Dec. 2013.

[95] A. To, "Video data visualization system: Semantic classification and personalization," *Int. J. Comput. Graph. Animation*, vol. 2, no. 1, pp. 1–64, 2008.

[96] Z. Wang et al., "Visual exploration of sparse traffic trajectory data," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1813–1822, Dec. 2014.

[97] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal, "Harnessing object and scene semantics for large-scale video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3112–3121.

[98] E. J. Soure, E. Kuang, M. Fan, and J. Zhao, "CoUX: Collaborative visual analysis of think-aloud usability test videos for digital interfaces," *Proc. IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 643–653, 2022.

[99] Y. Wu et al., "LiveRetro: Visual analytics for strategic retrospect in livestream e-commerce," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 1, pp. 1117–1127, Jan. 2024.

[100] J. He et al., "VideoPro: A visual analytics approach for interactive video programming," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 1, pp. 87–97, Jan. 2024.

[101] K. Maher et al., "E-ffective: A visual analytic system for exploring the emotion and effectiveness of inspirational speeches," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 508–517, Jan. 2022.

[102] Q. Chen et al., "ViSeq: Visual analytics of learning sequence in massive open online courses," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 3, pp. 1622–1636, Mar. 2020.

[103] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. B. I. Badia, "A human-AI collaborative approach for clinical decision making on rehabilitation assessment," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–14.

[104] T. Polk, D. Jäckle, J. Häußler, and J. Yang, "CourtTime: Generating actionable insights into tennis matches using visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 397–406, Jan. 2020.

[105] J. Lan, J. Wang, X. Shu, Z. Zhou, H. Zhang, and Y. Wu, "RallyComparator: Visual comparison of the multivariate and spatial stroke sequence in table tennis rally," *J. Visualizat.*, vol. 25, no. 1, pp. 143–158, Feb. 2022, doi: 10.1007/s12650-021-00772-0.

[106] D. Seebacher, T. Polk, H. Janetzko, D. A. Keim, T. Schreck, and M. Stein, "Investigating the sketchplan: A novel way of identifying tactical behavior in massive soccer datasets," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 4, pp. 1920–1936, Apr. 2023.

[107] L. Clarke, E. Hornecker, and I. Ruthven, "Fighting fires and powering steam locomotives: Distribution of control and its role in social interaction at tangible interactive museum exhibits," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–17.

[108] H. Guo et al., "DanceVis: Toward better understanding of online cheer and dance training," *J. Visualizat.*, vol. 25, no. 1, pp. 159–174, Feb. 2022.

[109] X. Xie et al., "PassVizor: Toward better understanding of the dynamics of soccer passes," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1322–1331, Feb. 2021.

[110] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4975–4986.

[111] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1273–1283.

[112] W. Luo, W. Liu, D. Lian, and S. Gao, "Future frame prediction network for video anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7505–7520, Nov. 2022.

[113] N. V. Laptev, V. V. Laptev, O. M. Gerget, A. A. Kravchenko, and D. Y. Kolpashchikov, "Visualization system for fire detection in the video sequences," *Sci. Vis.*, vol. 13, no. 2, pp. 1–9, 2021.

[114] G. Sun et al., "VSumVis: Interactive visual understanding and diagnosis of video summarization model," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 4, pp. 1–28, Aug. 2021.

[115] M. Fan, K. Wu, J. Zhao, Y. Li, W. Wei, and K. N. Truong, "VisTA: Integrating machine intelligence with visualization to support the investigation of think-aloud sessions," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 343–352, Jan. 2020.

[116] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu, "M2Lens: Visualizing and explaining multimodal models for sentiment analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 802–812, Jan. 2022.

[117] H. Liao, L. Chen, Y. Song, and H. Ming, "Visualization-based active learning for video annotation," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2196–2205, Nov. 2016.

[118] D. Connaghan, P. Kelly, and N. E. O'Connor, "Game, shot and match: Event-based indexing of tennis," in *Proc. 9th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2011, pp. 97–102.

[119] J. Piazentin Ono, A. Gjoka, J. Salamon, C. Dietrich, and C. T. Silva, "HistoryTracker: Minimizing human interactions in baseball game annotation," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–12.

[120] T. Zhang, A. El Ali, C. Wang, A. Hanjalic, and P. Cesar, "RCEA: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–15.

[121] A. Baldacci, F. Ganovelli, M. Corsini, and R. Scopigno, "Presentation of 3D scenes through video example," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 9, pp. 2096–2107, Sep. 2017.

[122] P.-Y. Chi, J. Liu, J. Linder, M. Dontcheva, W. Li, and B. Hartmann, "DemoCut: Generating concise instructional videos for physical demonstrations," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2013, pp. 141–150.

[123] P. O'Donovan and A. Hertzmann, "AniPaint: Interactive painterly animation from video," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 3, pp. 475–487, Mar. 2012.

[124] C. Lu, Y. Xiao, and C.-K. Tang, "Real-time video stylization using object flows," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 6, pp. 2051–2063, Jun. 2018.

[125] M. Liao, J. Gao, R. Yang, and M. Gong, "Video stereolization: Combining motion analysis with user interaction," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 7, pp. 1079–1088, Jul. 2012.

[126] Z. Chen et al., "Augmenting sports videos with VisCommentator," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 824–834, Jan. 2022, doi: 10.1109/TVCG.2021.3114806.

[127] J. J. Y. Chung, H. V. Shin, H. Xia, L.-Y. Wei, and R. H. Kazi, "Beyond show of hands: Engaging viewers via expressive and scalable visual communication in live streaming," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–14.

[128] D. Deng et al., "EventAnchor: Reducing human interactions in event annotation of racket sports videos," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–13.

[129] J.-C. Yoon, I.-K. Lee, and H. Kang, "Video painting based on a stabilized time-varying flow field," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 1, pp. 58–67, Jan. 2012.

[130] M. Flagg and J. M. Rehg, "Video-based crowd synthesis," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 11, pp. 1935–1947, Nov. 2013.

[131] S. Walton, M. Chen, and D. Ebert, "Live layer live traffic projection onto maps," in *Proc. Eurographics*, Jul. 2011, pp. 1–23.

[132] J. Schöning, P. Faion, G. Heidemann, and U. Krumnack, "Providing video annotations in multimedia containers for visualization and research," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 650–659.

[133] E. Wu, M. Piekenbrock, T. Nakumura, and H. Koike, "SPinPong–virtual reality table tennis skill acquisition using visual, haptic and temporal cues," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 5, pp. 2566–2576, May 2021.

[134] L. Turban, F. Urban, and P. Guillotel, "Extrafoveal video extension for an immersive viewing experience," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 5, pp. 1520–1533, May 2017.

[135] G. Andrienko et al., "Constructing spaces and times for tactical analysis in football," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 4, pp. 2280–2297, Apr. 2021.

[136] M. Stein et al., "Bring it to the pitch: Combining video and movement data to enhance team sport analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 13–22, Jan. 2018.

[137] X. Chu et al., "TIVEE: Visual exploration and explanation of badminton tactics in immersive visualizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 118–128, Aug. 2022.

[138] A. Serrano et al., "Motion parallax for 360° RGBD video," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 5, pp. 1817–1827, May 2019.

[139] Y. Liu, Q. Dai, and W. Xu, "A point-cloud-based multiview stereo algorithm for free-viewpoint video," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 3, pp. 407–418, May 2010.

[140] S.-S. Lin, C.-H. Lin, I.-C. Yeh, S.-H. Chang, C.-K. Yeh, and T.-Y. Lee, "Content-aware video retargeting using object-preserving warping," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 10, pp. 1677–1686, Oct. 2013.

[141] T. Tang, J. Tang, J. Hong, L. Yu, P. Ren, and Y. Wu, "Design guidelines for augmenting short-form videos using animated data visualizations," *J. Visualizat.*, vol. 23, no. 4, pp. 707–720, Aug. 2020.

[142] R. Currano, S. Y. Park, D. J. Moore, K. Lyons, and D. Sirkin, "Little road driving HUD: Heads-up display complexity influences drivers' perceptions of automated vehicles," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–15.

[143] T. A. Biresaw, T. Nawaz, J. Ferryman, and A. I. Dell, "ViTBAT: Video tracking and behavior annotation tool," in *Proc. 13th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2016, pp. 295–301.

[144] T. Löwe, M. Stengel, E. C. Förster, S. Grogorick, and M. Magnor, "Gaze visualization for immersive video," in *Eye Tracking and Visualization: Foundations, Techniques, and Applications*. Cham, Switzerland: Springer, 2017, pp. 57–71.

[145] J. Zhang, E. Langbehn, D. Krupke, N. Katzakis, and F. Steinicke, "Detection thresholds for rotation and translation gains in 360° video-based telepresence systems," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 4, pp. 1671–1680, Apr. 2018.

[146] C. Lee et al., "A visual analytics system for exploring, monitoring, and forecasting road traffic congestion," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 11, pp. 3133–3146, Nov. 2020.

[147] S. Ye et al., "ShuttleSpace: Exploring and analyzing movement trajectory in immersive visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 860–869, Feb. 2021.

[148] Y. Wu et al., "ForVizor: Visualizing spatio-temporal team formations in soccer," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 65–75, Jan. 2019.

[149] W. C. Payne et al., "DanceON: Culturally responsive creative computing," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–16.

[150] O. Kaplan, G. Yamamoto, Y. Yoshitake, T. Taketomi, C. Sandor, and H. Kato, "In-situ visualization of pedaling forces on cycling training videos," in *Proc. IEEE Int. Conf. Syst. Man, Cybern.*, Oct. 2016, pp. 994–999.

[151] M. L. Parry, P. A. Legg, D. H. S. Chung, I. W. Griffiths, and M. Chen, "Hierarchical event selection for video storyboards with a case study on snooker video visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 1747–1756, Dec. 2011.

[152] C. Dietrich, D. Koop, H. T. Vo, and C. T. Silva, "Baseball4D: A tool for baseball game reconstruction & visualization," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2014, pp. 23–32.

[153] H.-T. Chen, W.-J. Tsai, and S.-Y. Lee, "Contour-based strike zone shaping and visualization in broadcast baseball video: Providing reference for pitch location positioning and strike/ball judgment," *Multimedia Tools Appl.*, vol. 47, no. 2, pp. 239–255, Apr. 2010.

[154] G. Halter, R. Ballester-Ripoll, B. Flueckiger, and R. Pajarola, "VIAN: A visual annotation tool for film analysis," *Comput. Graph. Forum*, vol. 38, no. 3, pp. 119–129, Jun. 2019.

[155] Q. Chen, Y. Chen, D. Liu, C. Shi, Y. Wu, and H. Qu, "PeakVizor: Visual analytics of peaks in video clickstreams from massive open online courses," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 10, pp. 2315–2330, Oct. 2016.

[156] C. Shi, S. Fu, Q. Chen, and H. Qu, "VisMOOC: Visualizing video clickstream data from massive open online courses," in *Proc. IEEE Conf. Vis. Anal. Sci. Technol. (VAST)*, Oct. 2014, pp. 277–278.

[157] G. Zhang, Z. Zhu, S. Zhu, R. Liang, and G. Sun, "Towards a better understanding of the role of visualization in online learning: A review," *Vis. Informat.*, vol. 6, no. 4, pp. 22–33, Dec. 2022, doi: 10.1016/j.visinf.2022.09.002.

[158] Y. Wang, Y. Liu, X. Tong, Q. Dai, and P. Tan, "Outdoor markerless motion capture with sparse handheld video cameras," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 5, pp. 1856–1866, May 2018.

[159] L. Pan et al., "Visual interactive image clustering: A target-independent approach for configuration optimization in machine vision measurement configuration optimization in machine vision measurement," *Frontiers Inf. Technol. Electron. Eng.*, vol. 24, no. 3, pp. 355–372, Mar. 2023.

[160] Z. Feng, Z. Zeng, C. Guo, and Z. Li, "Temporal multimodal graph transformer with global-local alignment for video-text retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1438–1453, Mar. 2023.

[161] J. Dong et al., "Reading-strategy inspired visual representation learning for text-to-video retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5680–5694, Aug. 2022.

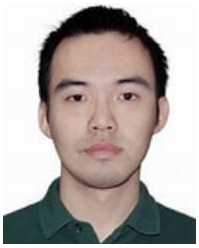
[162] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[163] B. Wu, G. Niu, J. Yu, X. Xiao, J. Zhang, and H. Wu, "Towards knowledge-aware video captioning via transitive visual relationship detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6753–6765, Oct. 2022.

[164] J. Zhu, P. Zeng, L. Gao, G. Li, D. Liao, and J. Song, "Complementarity-aware space learning for video-text retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 1–14, Dec. 2023.

[165] L. Gao, Y. Lei, P. Zeng, J. Song, M. Wang, and H. T. Shen, "Hierarchical representation network with auxiliary tasks for video captioning and video question answering," *IEEE Trans. Image Process.*, vol. 31, pp. 202–215, 2022.

**Wang Xia** received the B.E. degree in software engineering from the College of Computer Science and Technology, Zhejiang University of Technology, in 2019, where she is currently pursuing the Ph.D. degree in computer science and technology. Her main research interests include interactive machine learning, image/video analysis, and information visualization.



**Guodao Sun** received the B.Sc. degree in computer science and technology and the Ph.D. degree in control science and engineering from Zhejiang University of Technology, Hangzhou, China. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology. His main research interests include urban visualization, visual analytics of social media, and information visualization.

**Tong Li** is currently pursuing the Ph.D. degree in computer science and technology with the College of Computer Science and Technology, Zhejiang University of Technology. Her main research interests include information visualization, video visualization, and computer vision.

**Baofeng Chang** received the B.Sc. degree in automation and the Ph.D. degree in computer science and technology from Zhejiang University of Technology. He is currently with Zhejiang Airport Innovation Institute, Zhejiang Provincial Airport Group, and Zhejiang University of Technology, China. His main research interests include dynamic network visualization, traffic information visualization, and temporal sequence visualization.

**Jingwei Tang** received the B.E. degree in communication engineering from Zhejiang University of Technology, Hangzhou, China, in 2017, where he is currently pursuing the Ph.D. degree in computer science and technology. His research interests include data mining, visual analytics of network data, and information visualization.

**Gefei Zhang** received the B.Ed. degree in education technology from the College of Educational Science and Technology, Zhejiang University of Technology, Hangzhou, China. She is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University of Technology. Her main research interests include information visualization and educational data mining.

**Ronghua Liang** received the Ph.D. degree in computer science from Zhejiang University. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology, China. His research interests include visual analytics and computer vision.