LANDER: Visual Analysis of Activity and Uncertainty in Surveillance Video

Tong Li, Guodao Sun✉, Baofeng Chang, Yunchao Wang, Qi Jiang, Yuanzhong Ying, Li Jiang, Haixia Wang, Ronghua Liang

***Abstract*—Vision algorithms face challenges of limited visual presentation and unreliability in pedestrian activity assessment. We introduce LANDER, an interactive analysis system for visual exploration of pedestrian activity and uncertainty in surveillance videos. This visual analytics system focuses on three common categories of uncertainties in object tracking and action recognition. LANDER offers an overview visualization of activity and uncertainty, along with spatio-temporal exploration views closely associated with the scene. Expert evaluation and user study indicate that LANDER outperforms traditional video exploration in data presentation and analysis workflow. Specifically, compared to the baseline method, it excels in reducing retrieval time (p<0.01), enhancing uncertainty identification(p<0.05), and improving the user experience (p<0.05).**

***Index Terms*—Visual analysis, surveillance video, spatio-temporal activity, uncertainty.**

I. Introduction

PEDESTRIAN activity in surveillance videos encompass multi-dimensional information like spatio-temporal trajectories, action details, and scene context. While analyzing pedestrian activity is crucial for understanding and researching surveillance video, the process is typically time-consuming [1], [2]. Recently, the application of vision technology has significantly reduced the workload of security personnel in certain aspects [3]. But when it comes to surveillance video analysis in real-world scenarios, limitations still exist: vision algorithms mostly output image sets or video clips, which still rely on traditional retrieval mode. Such limited visual presentation and interaction may hinder in-depth analysis of complex pedestrian activity [4]. Also, automatic video analysis technologies are not entirely reliable for analyzing pedestrian activity in surveillance video [5]. Data quality and "black box" models can lead to uncertainties in trajectory, action, or activity recognition [6].

Experts interview reveals an interesting phenomenon: security personnel utilize intelligent technology to quickly locate targets from numerous cameras in real-time. Yet, for specific video analysis tasks, they still rely on manually viewing the entire video. They seek *"innovative tools to reduce workload and alleviate concerns"*. The focus of visualization researchers on "human-in-the-loop" offers novel insights into video analysis. To alleviate limited visual representation in video retrieval, some studies [7]–[9] integrated trajectory and action data into image and video using elements like points, lines, and glyphs. In response to uncertainty, some studies [5], [6] focused on implicit uncertainty in feature processing. However, this implies an additional burden for security personnel, as they need to acquire domain knowledge to interpret uncertainties. In contrast, we focus on common uncertainties in automatic trajectory and action recognition. This is more intuitive and understandable for users lacking technical background, eliminating the need to focus on algorithmic details. Additionally, while data source-embedded visualizations offer useful visual summaries, their lack of interaction limits ability to query and explore complex activity data. We developed LANDER, a visual analysis system. It emphasizes interactive exploration of spatio-temporal patterns in trajectory, action, and uncertainty, advancing beyond the basic level of just "viewing" data.

We proposed an uncertainty analysis method based on covariance matrices and self-information. It quantifies three abstract uncertainties in automatic trajectory and action recognition. Compared to explainable AI methods focused on hidden features, our approach is more accessible to users without domain-specific knowledge. Additionally, it is not limited to specific video sources or vision models. We designed an interactive visual analysis system LANDER. It embeds visualization elements into the context of pedestrian activity in video source, enabling analysts to track and explore pedestrian activity and uncertainty during video retrieval. In addition, LANDER utilizes a hierarchical view design, supporting the exploration of spatio-temporal patterns from overview to detail. Specifically, there is an about 58.14% improvement in user experience. We evaluated LANDER across multiple real-world surveillance video scenarios. Evaluation results demonstrate that LANDER reduces video retrieval time by approximately 22.91%, lowers the uncertainty identification error rate by about 58.20%. And most improvements achieved statistical significance. Participants indicated that, compared to manually searching the entire video, they preferred to use the LANDER.

In summary, the contributions of this paper are as follows:

- An explicit uncertainty quantification method based on covariance matrix and self-information, not limited to specific video sources and vision models.
- An interactive visual analysis tool anchors pedestrian activity and uncertainty visualizations within scene context, distinct from mere visual summary tools.
- An evaluation in multiple real surveillance scenarios. The results show a 22.91% reduction in search time and a 58.20% decrease in identification errors, with most enhancements statistically significant.

T. Li, G. Sun, B. Chang, Y. Wang, Q. Jiang, Y. Ying, L. Jiang and H. Wang are with the College of Computer Science and Technology, Zhejiang University of Technology. R. Liang is with Zhejiang University of Science and Technology. (e-mail: litong@zjut.edu.cn, guodao@zjut.edu.cn, baofeng.chang@foxmail.com, wyctears@gmail.com, jiangqi@zjut.edu.cn, yuanzhongying@zjut.edu.cn, jl@zjut.edu.cn, hxwang@zjut.edu.cn, rhliang@zjut.edu.cn)

Manuscript received April 00, 0000; revised August 00, 0000.

II. Related Work

A. Vision-based Video Understanding Techniques

Research centered on Siamese Network marks a significant breakthrough of deep learning in object tracking. Beginning with [10], this series progressively enhanced the accuracy of tracking frames [11] and the fineness of object representation [12]. Recent studies such as RPformer [13], Repformer [14] and EANTrack [15] have emphasized the importance of global contextual information. These innovative network structures and attention mechanisms demonstrate significant tracking ability in complex scenarios with scale changes, occlusions, and rapid movements. Research in multi-object tracking focuses on enhancing tracking accuracy and stability by modeling the spatio-temporal relationships between objects. Wojke et al. [16] followed the framework of [17] but introduced appearance features and matching cascade to enhance object differentiation. Recent studies [18], [19] enhance tracking performance in non-linear motions and extreme occlusions by improving traditional algorithms and introducing novel motion learning mechanisms.

In video action recognition, techniques progressed from initial Two-stream Network [20] to 3D Convolutional Neural Network [21] for better spatio-temporal feature capture. Subsequently, the application of Long short-term memory networks further enhanced the handling of temporal features [22]. Recent studies [23], [24] have significantly improved action recognition outcomes through self-supervised and contrastive learning, even with limited annotated data. Gao et al. [25] innovatively introduced the multitemporal scale and spatio-temporal transformer network, effectively optimizing action boundaries and classification.

Vision-based video techniques excel in efficiently processing large volumes of video data and extracting key information. However, their outputs are often presented as discrete image sets and video clips. Such limited visual presentation and interaction may hinder users' deep analysis of complex pedestrian activity [4]. We introduce the interactive visual analysis approach to alleviate this shortfall. We transform activity data into intuitive visual elements, enhancing the recognition of pedestrian activity patterns. Interactive exploration alters traditional video retrieval, offering flexible and structured data querying and analysis approach.

B. Visualization-focused Surveillance Video Analysis

Researchers have applied visualization in various scenarios, providing fresh insights for analyzing diverse video types. Relevant research includes traffic monitoring, medical diagnosis, educational understanding, sports performance analysis, as well as e-commerce marketing, among others [26]–[36].

Trajectory mining and summarization is a key research direction in surveillance video visualization. Researchers utilized trajectory lines and heatmaps for visualizing object movement paths and activity intensity [5], [7], [37]–[39]. These visualization schemes enable a more intuitive and clear understanding of activity patterns and anomalies. Activity monitoring aims to parse essential information from behavior and events, offering users comprehensive visual insights. The 3D cube visualization is an effective three-dimensional representation technique. By combining two spatial dimensions with time or other metrics, it effectively displays data's dynamic spatio-temporal distribution and interrelations [40]–[42]. Augmented visualization embedded in data sources (image, video, 3D space) integrates data with the scene for a realistic and immersive visualization [5], [9], [35].

Works closely related to our research include [7] and [6]. The former introduced the VideoPerpetuoGram (VPG), an ECG-like video stream visualization technology. This allows for a condensed, multi-dimensional data overview within the video sequence. The latter integrated VPG into their visual analytics framework, effectively combining automatic video analysis with interactive feature filters. This approach demonstrated significant advantages in analysis efficiency for large-scale video data. However, the VPG method constrains the potential for in-depth data exploration and interaction. Our approach emphasizes user-focused and interactive exploration of pedestrian spatio-temporal activity patterns, rather than mere "watching". Furthermore, the former utilized action prediction plausibility to quantify uncertainty, aiding intuitive understanding of object actions and relations. In contrast, the latter focused on implicit uncertainties in feature extraction. Our research focuses on summarizing and presenting three explicit uncertainties in automatic activity assessment results. It is not limited to specific video sources or vision models. This is intuitive for users without domain-specific knowledge, as it does not require attention to algorithmic details.

III. Formative Study and User-Centered Design

We adopt a user-centered approach for investigation and design. Initially, we explored research in surveillance video visual analysis to identify limitations of current technologies and unmet needs. Subsequently, we conducted a face-to-face interview with experts in video surveillance field [43]. The goal was to gain practical experience and insights, understanding their daily work challenges and needs.

A. Literature Investigation

We interviewed a professor E0 (male, $age = 38$, $work = 14$) with extensive experience in data analysis and visualization. His expertise includes 8 years in interdisciplinary research combining computer vision and visual analysis, which is highly relevant to our study topic. Under E0's guidance, we identified three main objectives for the literature investigation: **LG1** what attributes are focused on? Surveillance videos encompass data ranging from individual characteristics to group relations. This objective aims to identify the attributes commonly emphasized in research. **LG2** how visualizations are designed and presented? Appropriate visual representation is crucial for revealing complex video data. This objective aims to understand the visualization practices in research. **LG3** how uncertainties are considered? Uncertainty introduced by vision processing may impact decision-making. This objective aims to understand the discussion of uncertainty in research.

Here, we listed 19 typical studies in surveillance video visual analysis and summarized them across six key dimensions: **(LG1)** Attribute **(LG2)** Visual Design, Visual Medium

(LG3) Vision Techniques, Uncertainty Quantification and Uncertainty Visualization, as shown in Fig. 1. *Please refer to supplementary material Section I for details.* **In conclusion:** (1) Advancements in neural network-based techniques have spurred increased focus on biometric feature research. Nevertheless, the analysis of trajectory (42.10%) and coarse-grained activity (36.86%) continue to dominate the field. (2) Although surveillance video data is often complex and information-rich, our observation reveals that most research still tends to use basic and intuitive design schemes. (3) Current research often focuses on a single visualization medium, with a high prevalence of data source embedding (52.63%). These efforts typically emphasize visual summary and presentation but overlook the crucial role of interactive exploration in video surveillance analysis. (4) Current studies lack focus on uncertainty introduced by vision processing techniques. Undeniably, identifying, quantifying (26.32%), and effectively visualizing (21.05%) this uncertainty is crucial for ensuring the accuracy and reliability of analysis results.

* **LG1** What attributes are focused on? **LG2** How visualizations are designed and presented? **LG3** How uncertainties are considered?

LG1			LG2								LG3	
				Visual Medium					Vision Techniques			
	Attribute	Literatures	Visual Design	Data Source	User Interface	Flow Field	Illustration		ANN-based	TIP-based	Uncertainty Quantification	Uncertainty Visualization
10.52%	Biometric Features	[51]	line/bar		■				■		"confidence"	text
		[52]	line/bar		■				■		-	-
42.10%	Trajectory	[8]	VPG	■						■	-	-
		[9]	line	■						■	-	-
		[6]	VPG	■					■		"error propagation"	blur
		[40]	3D cube	■						■	-	-
		[37]	line				■			■	-	-
		[5]	VPG	■					■		"error propagation"	blur
		[38]	line/heatmap	■					■		-	-
		[39]	line	■					■		"error"	-
10.52%	Action	[7]	VPG	■						■	"plausibility"	thickness/saturation
		[40]	Image shot	■						■	-	-
36.86%	Activity	[53]	flow visualization			■				■	-	-
		[54]	flow visualization			■				■	-	-
		[41]	3D cub/heatmap		■					■	-	-
		[42]	3D cub/heatmap		■					■	-	-
		[55]	line/area	■						■	-	-
		[44]	component		■				■		-	-
		[56]	line/bar		■				■		-	-
				52.63%	31.58%	10.52%	5.27%				26.32%	21.05%

***Visual Medium** refers to four types of mediums used: data source(like images, videos, 3D models), user interfaces, flow fields, and illustrations.
Vision Techniques summarizes the vision processing techniques used in video analysis flow, including ANN-based and TIP-based.
ANN: Artificial Neural Network TIP: Traditional Image Processing

Fig. 1. Literature investigation statistics: concentrates on data attributes, visual design, and uncertainty discussions.

B. Experts Interview

We invited three experts (E1, E2 and E3) for a 2-hour face-to-face interview. E1 (male, $age = 36$, $work = 8$) is an associate professor at a police academy. He specializes in visual analysis of police data, and is able to provide essential guidance for LANDER's design. E2 (male, $age = 41$, $work = 16$) is the head of technical department at a public security bureau. His insights can assist us in understanding the challenges of automated technology in applications, ensuring LANDER's technological advancement. E3 (male, $age = 32$, $work = 6$), head of the video surveillance team, can offer practical advice for LANDER based on his familiarity with video surveillance tools and procedures. The interview focused on three topics: **EG1** which pedestrian activities are generally the focus? This topic aims to identify primary tasks for pedestrian activity retrieval, clarifying focus data and tasks of this paper. **EG2** what tools are commonly used for retrieving surveillance video? This topic explores the limitations of common retrieval tools and considers how visualization techniques can help improve them. **EG3** what are the attitudes towards automatic assessment technologies? This topic aims to understand practical shortcomings, considering how visualization can bridge these gaps.

C. Experts Feedback and Task Analysis

We synthesized interview feedback with literature review, identifying key tasks in this paper from three levels:

Supporting visual exploration of spatio-temporal activity. Activity-level tasks focus on enhancing visual representation of pedestrian activity, reducing the workload of security personnel in handling long videos.

T1 Presenting the spatio-temporal evolution of pedestrian activity is a critical task. Experts hope to quickly grasp activity's context through an overview view (how). When discussing video retrieval tools, E3 expressed the desire for visualization:*"I usually operate with a video player, but often forget the content. I would like visualization to assist, particularly with complex trajectories and frequent actions change."* E2 added, *"Especially when longer video need to be retrieved."* Visualization researchers also focus on summarizing trajectories visually and are dedicated to supporting pattern analysis in trajectory and action details [7], [37], [40].

T2 Identifying pedestrian activity in specific locations and times is an important focus task. This enables security personnel to swiftly pinpoint activities from temporal and spatial dimensions (when and where). E2 mentioned the frequent need to search specific activity:*"Operators usually focus on key places or specific times in a case, rather than everything."* Visualization tools are often designed for exploring specific location and time information to reveal finer-grained activity patterns of individuals or groups [49], [51].

Supporting the recognition and visual exploration of uncertainty in spatio-temporal pedestrian activity. Uncertainty-level tasks focus on guiding security personnel to identify potential uncertainties in automated activity assessments, aiming to reduce misinformation.

T3 Revealing potential uncertainties to boost security personnel's confidence in using intelligent algorithms for auxiliary review tasks (which). When discussing automatic vision techniques, both E2 and E3 mentioned unexpected recognition scenarios that were often not communicated in advance. The frequently mentioned issues are *"problems with tracking a person"* and *"either incorrect or missed detections in actions"*. As stated in Section III-A, most video visualization efforts have not focused on potential impacts, nor have they emphasized using visualization to fill this gap.

T4 Modeling the abstract uncertainty, enabling security personnel lacking domain knowledge to understand uncertainty degree through numbers and visual designs (what). E1 suggested:*"Intuitive numbers and visualizations instead of vision technicians to alert security personnel to the presence and degree of uncertainty."* E3 added, *"I particularly don't want to focus on vision technology details; perhaps numbers, colors, and graphics would be more helpful for me."*

T5 Presenting the spatio-temporal changes of uncertainty along with pedestrian activity is also necessary (how). E1 mentioned, *"If uncertainty information is visualized and combined with pedestrian activity, it may be more intuitive!"* E2 expressed agreement with this viewpoint.

Supporting scene verification. Scene-level tasks aim to enhance factual validation for security personnel in the decision-making process.

T6 Associating data with scene context. The visualization of pedestrian activities and their uncertainties should be closely integrated with the video scenes (how). This allows security personnel to observe the factual context while exploring the visualizations. E1 mentioned, *"Based on my observations, security personnel often prefer immersive analysis without disconnecting from the video."* E2 noted, *"Visualization is an auxiliary tool, and I will ultimately make decisions based on the video itself. This is my professional habit and integrity."* Most visual research related to surveillance activities considers associating visual design with scenes, especially when it involves pedestrian trajectory analysis [7], [9], [38].

Additionally, E2 expressed a desire for the tool to save important video segments for easier reporting and collaboration. We also take this request into consideration. It is worth noting that other design considerations were involved in the discussions with experts. *Please refer to supplementary material Section II for details.*

D. Pipeline

As shown in Fig. 2, we construct the pipeline for visual exploration of uncertainty-based spatio-temporal activity: (A) Activity data processing module includes spatio-temporal activity data acquisition based on computer vision algorithms. (B) Uncertainty analysis module includes pedestrian activity uncertainty definition and mathematical modeling. (C) Visual design module mainly includes visual design and concise interaction based on information fusion, detailed information focus and scene context awareness.

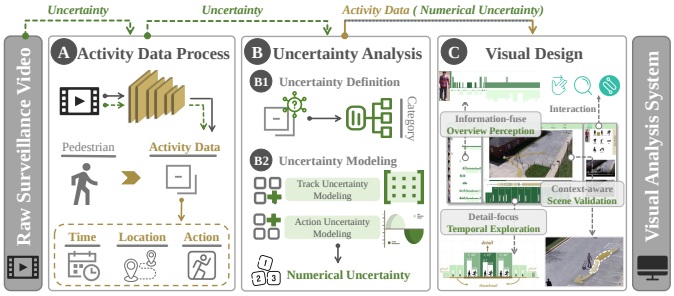


Fig. 2. Visual analysis pipeline for pedestrian activity and three categories of uncertainty.

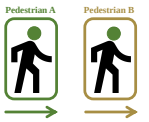
IV. Uncertainty Concept and Analysis

We focus on activity analysis in two key dimensions: trajectory and action, along with their uncertainty analysis and modeling. For details on vision model deployments, apparatus, and the general structure of activity data, *please refer to supplementary material Section III*.

However, automation algorithms may introduce uncertainties in trajectory, action, or activity recognition. Next, we will discuss certain categories of uncertainties and their modeling process, aiming to quantitatively express uncertainties in spatio-temporal activity data.

A. Uncertainty Category

Activity uncertainties introduced by automation algorithms are complex, multidimensional, and abstract. We reached consensus with domain experts, focusing on three main uncertainties in the aforementioned two vision tasks. These categories represent common challenges faced by experts, reflecting the general limitations of vision algorithms in applications (**T3**).



Category A: Misidentification Uncertainty. Due to limitations in object identification and tracking, the algorithm may incorrectly recognize *Pedestrian A* as a new *Pedestrian B*. This can affect understanding of the object's complete activities.

Category B: Tracking Missing Uncertainty. Instabilities in the tracking algorithm can result in incomplete tracking of *Pedestrian A*, leading to temporary loss of tracking. This can impact awareness of the object's activities during these intervals.

Category C: Action Recognition Uncertainty. With algorithms' limited capability or sensitivity to subtle actions, predicted actions may not align with *Pedestrian A*'s actual state. This can lead to potential misinterpretations of the object's actions.

B. Covariance Matrix-based Tracking Uncertainty Modeling

The aforementioned categories of uncertainties present challenges in activity analysis. To mitigate this, we conduct mathematical modeling to facilitate subsequent analysis and visualization efforts (**T4**).

Category A primarily arises from the tracking algorithm's limited capability to consistently recognize the same pedestrian. The degree of this category can be assessed based on the similarity of pedestrians' characteristics. In this paper, however, we do not include this module. Instead, we explore preliminary the visualization-based solution in Section V-A. For **Category B**, uncertainty primarily arises from unknown pedestrian trajectories and actions during temporary tracking missing, i.e. missing data.

Justification: In statistics, fixed values or summary statistics (such as mean, median, mode) are commonly used to fill missing data. Or, incomplete data entries are discarded. However, these methods are unsuitable for video contexts emphasizing dynamism and temporality, potentially leading to contextual errors or loss. Therefore, we consider a method capable of estimating pedestrian states and assessing their uncertainties. The Kalman filter [52], suitable for dynamic systems and widely used in object tracking, reliably estimates current states based on previous ones. It iteratively refines estimates and covariance matrices for uncertainty assessment. We do not focus on optimal state estimation using a Kalman filter. Rather, our emphasis is on quantifying **Category B** through estimation errors. This involves three stages:

Stage 1: Estimation Assume that the tracking missing period is $[t_i, t_j]$. $trajectory_{t_{i-1}}$ and $trajectory_{t_{j+1}}$ represent the trajectories at t_{i-1} and t_{j+1}, respectively. We formulate a state equation and construct a filter (*details in supplementary*

material Section IV). This filter estimates missing trajectories within $[t_i, t_j]$, starting with $trajectory_{t_{i-1}}$ and $trajectory_{t_{j+1}}$ as the initial input, as shown in Fig. 3(1). This estimation proceeds until the trajectory error minimizes and aligns closely with $trajectory_{t_{j+1}}$. At each step of estimation, we derive an estimated trajectory and a covariance matrix P. The estimated trajectory indicates the probable location during the missing period, while the P quantifies the associated uncertainty.

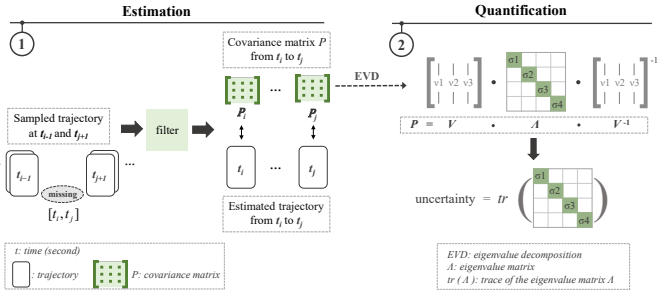


Fig. 3. **Category B** measurement: covariance matrix-based measurement for tracking missing uncertainty.

Stage 2: Quantification For analysts, understanding the covariance matrix P can be challenging compared to a numerical format and visualization. We therefore consider converting it into a numerical value for better clarity. The eigenvalues of covariance matrix indicate the dispersion degree. Thus, we apply eigenvalue decomposition (EVD) on P to obtain the diagonal matrix Λ, as shown in Fig. 3(2). Further, the uncertainty degree of estimated trajectory is quantified by calculating the trace, $tr(\Lambda)$.

Stage 3: Normalization Additionally, we assigned uncertainty values to sampled trajectories obtained by stable tracking algorithms. This is because, despite being stably tracked, these trajectories may still hold slight uncertainties due to inherent variations in neural networks. Finally, we normalize these uncertainty values using min-max normalization.

C. Self Information-based Action Uncertainty Modeling

Action recognition algorithms typically produce multiple action labels with corresponding confidence levels. Confidence reflects the probability of an action's occurrence, offering an intuitive way to quantify uncertainty. For **Category C**, we apply a quantification strategy based on confidence and self-information to measure the uncertainty.

Justification: In single-action predictions, lower confidence indicates increased uncertainty; conversely, decreased uncertainty. In multi-action predictions, smaller differences in confidences between similar actions imply higher uncertainty; conversely, the lower uncertainty. For example, predictions like $[stand, 0.88]$ and $[walk, 0.87]$ indicate ambiguity due to their closely matched confidence scores. This aligns with the concept in information theory linking self-information to uncertainty: $I(p_i) = -log(p_i)$, where I denotes self-information and p_i represents the probability of event.

In our experimental observations, we noted that action recognition algorithms are not sufficiently sensitive to minor changes in lower limb movements in surveillance environments. This often leads to ambiguous predictions for actions like standing and walking. Therefore, when similar lower limb actions occur concurrently, uncertainty is measured by the differences in their confidence levels, as shown in Fig. 4. In other cases, it's measured by the confidence level itself. Given the singularity of logarithmic function at $c\_diff = 0$, we assign a fixed value to the measurement there. Ultimately, we apply min-max normalization to action uncertainty values under different conditions.

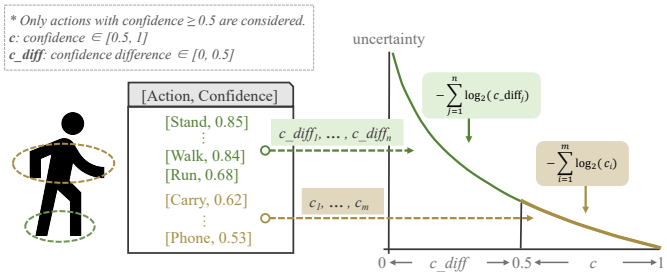


Fig. 4. **Category C** measurement: combining prediction confidence and self-information to measure action uncertainty.

V. Visual Design for LANDER

A. Information-fused Overview Perception

The surveillance scene involves numerous pedestrians and abundant activity information, making it challenging to obtain an overall overview of pedestrian spatio-temporal activity. The design goal is to employ visualization techniques to summarize, present, and compare multidimensional overview information of pedestrians. This aims to guide users to focus on key elements and important features within spatio-temporal activity and explore meaningful details within them (**T1, T5**).

Justification: In computer vision, pedestrian appearance features are commonly used for re-identification [16]. Inspired by this, we initially designed appearance feature bar to imply individual identity (Fig. 5A). In other words, multiple appearances of a bar in the view indicate **Category A**. That is, the algorithm fails to correctly identify the same person. However, E2 stated that in actual scenarios, the same individual's characteristic may vary due to changes in surveillance perspectives or pedestrian appearances. *"In this design, it becomes necessary to determine whether the feature bars belong to the same pedestrian, which is labor-intensive."* He suggested replacing bars with intuitive pedestrian images.

Fig. 5. Design considered but not adopted: (A) An appearance feature bar characterizes a pedestrian; (B) TreeMap encodes the overview of pedestrian activities in a video.

Justification: Additionally, we have considered the treemap to present overview information about pedestrians in a video (Fig. 5B). The size of rectangle represents activity level (determined by activity area and duration). The color represents overall uncertainty. Advantage of this design lies in its compact layout, reducing wastage of pixel space. It lacks effectiveness in indicating temporal information, which hinders the ability to capture activity and uncertainty trends over time. However, temporal information serves as a crucial foundation for video analysis. Pixel bar designs typically consist of multiple bars, each spanning only a few pixels in width. It is particularly suitable for displaying two-dimensional or higher, and excels in representing high-density data. Experts also expressed that the design is intuitive and user-friendly.

As shown in Fig. 6A, based on video duration, pixel bars are displayed based on the temporal anchor of pedestrian activity. The vertical pixel space is divided into two channels: the upper chart's pixel bars encode action type count and uncertainty **Category C** (height and color, respectively), while the lower chart's pixel bars correspond to trajectory centrality and uncertainty **Category B** (height and color, respectively). Centrality here indicates the trajectory's proximity to the primary activity center in the scene. The dual-channel temporal pixel chart provides users with an overview of time-based changes in multidimensional activity data.

However, displaying large-scale temporal charts in limited visual space inevitably leads to significant compression of pixel bars. This results in high-frequency variations in height and color (Fig. 6B), which may distract users from focusing on key aspects of pedestrian activity. To alleviate this issue, we propose a "two-step" sequential subregion aggregation method that uniformly aggregates the height and color variations. First, we divided the original pixel bar into equal-length subregions. Then, within each subregion, we used the pixel height with the maximum proportion as unified height mapping for height aggregation. The same approach is applied to color aggregation. It is possible to moderately reduce the frequency of pixel height and color variations by this method. This can provide clearer visual effects, helping users to observe activity overview more effortlessly.

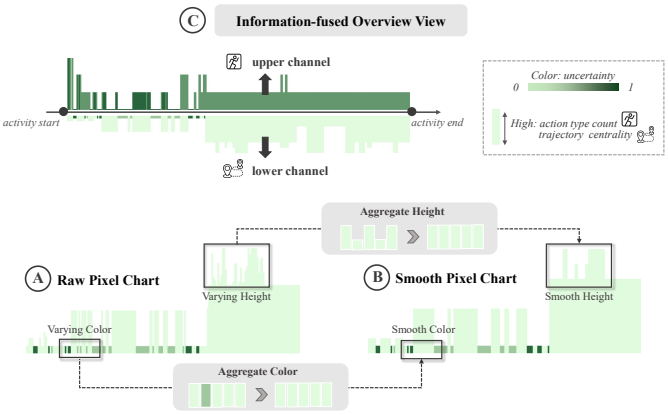


Fig. 6. (A) Dual-channel pixel chart encodes multidimensional information about a pedestrian's activity. (B) "Two-step" aggregation scheme combines frequently changing pixel heights and colors in subregions.

B. Detail-focused Temporal Exploration

The pixel bar charts provide an overview of pedestrian activity but faces challenges in exploring details. Therefore, this view is designed to support users in visually exploring activity details within specific periods (**T2**).

Video scene provides intuitive information on pedestrian trajectories and actions, offering high referential value. We reached an agreement with experts that separate visualizations for this information are unnecessary. However, experts have expressed concerns about the complexity (types and confidences) of actions in algorithmic prediction. Without visualizing these elements, it would be difficult to comprehend the reasons for **Category C**. E1: *"Although pedestrian actions can be observed in the video, the reasons for uncertainty seem difficult to discern."* We considered embedding relevant information into the video scene, but such a design would significantly obstruct the original scene. After consideration, we opted for a separate view to show temporal action details, and link them to video scene through interactions.

It is still difficult to show complete sequence information of actions with limited visual space. We considered the scale-transforming timeline (Fig. 7), which enables detailed magnification of focal area without occupying additional space. The near-focus context region is presented as thumbnails, while the far-focus context region remains compressed at original pixel scale. *Our considerations are as follows:* Firstly, when focusing on the local region, it is essential to maximize the detail display. Thus, these areas should occupy a predominant portion of the visual space. Secondly, pedestrian activities are strongly temporal. Presenting context in the near-focus aids users in understanding data through temporal connections. However, it is crucial to avoid distracting the visual focus. Hence only a moderate pixel space is dedicated to thumbnail information. Lastly, the far-focus context continues to be represented with smaller pixel size for overview.

The original timeline (Fig. 7B) only display information at the pixel level, while the scale-transforming timeline (Fig. 7C) present multi-scale temporal action information. As shown in Fig. 7A, "detail-level" information is presented in the focus region. **Category C** is encoded with rectangular color, and time label is positioned at the top. Bars inside the rectangle display action types and confidence levels. Icons symbolize the action types, and the bar height represents prediction confidence levels, accompanied by corresponding labels. In the near-focus region, time label and confidence label are removed, retaining "thumbnail-level" information, which includes bar's height and type icon. The far-focus context region maintains a "pixel-level" rectangular size. The rectangle's height encodes action type counts, while the color represents action uncertainty. The scale-transforming timeline provides multi-level information. It enables a focus on local details while preserving awareness of the overall temporal context.

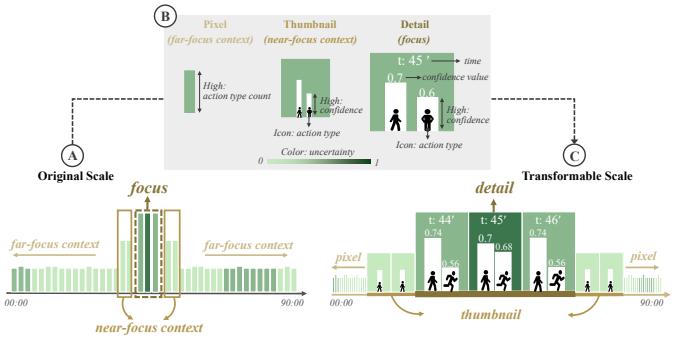


Fig. 7. Scale-transforming timeline design for actions details exploration. Different levels of visual encoding for focus, near-focus, and far-focus.

C. Context-aware Scene Verification

To accurately decide on pedestrian activity and uncertainty, security personnel must grasp the scene's context. Thus, it is essential to integrate activity and uncertainty data with the surveillance scene for context-aware scene verification. This enables users to intuitively validate pedestrian activity at specific times and locations (**T2, T6**).

Justification: Validation of the original video stream is essential. Therefore, we retained standard video player functions to align with security personnel's routine practices. Additionally, we positioned visual elements close to the activity context, facilitating rapid discernment of the connection between activity patterns and the scene. Given limited space within the video scene, we restricted information rendering to prevent overburdening the video review. We adopt an "on-demand rendering" approach for rendering, offering selection at sampling, activity, and uncertainty levels (Fig. 8).

The **Sampling Level** (Fig. 8A) only renders sampled trajectories, offering an overview of activity paths. In **Activity Level**, we introduced "activity bubble" to summarize the distribution of pedestrian lingering spaces and actions. *Justification:* Lingering activities often overlap in time while being spatially close, leading to frequent spatial intersections. This implies that spatial intersections frequently occur. Additionally, multiple actions commonly occur simultaneously, resulting in temporal intersections. Typically, stacked bar charts are utilized. However, such glyphs can obscure the video scene, thereby hindering video validation. To better understand the complex spatio-temporal relationships in pedestrian activity, we adopted the bubble-like visual encoding. This approach compactly encloses elements within a set and facilitates exploring inter-set relationships. Experts feedback was overwhelmingly positive: *"It is easy to understand and visually appears more refreshing."* As shown in Fig. 8B, "activity bubble" spatially encode the location of lingering activities or actions, while color encodes their duration.

The **Uncertainty Level** (Fig. 8C) renders "blur circle" and bounding box to represent the tracking uncertainty **Category B**. The estimated trajectory and error by Kalman filter (Section IV-C) can be used to describe the potential range and uncertainty level of missing trajectories. We used circles to encode spatial information of estimated trajectory. The center position, radius, and color correspond to the encoding of estimated trajectory's center, average radius, and uncertainty level. Blur is a suitable method for expressing uncertainty [53]. Inspired by this, we applied blur to the circle to represent uncertainty, calling it the "blur circle". Additionally, following E1's advice, we retained the bounding box in algorithm result to indicate pedestrian tracking status. By closely integrating the bounding box with scene context and combining it with the blur circle, it enhances users' understanding of **Category B**.

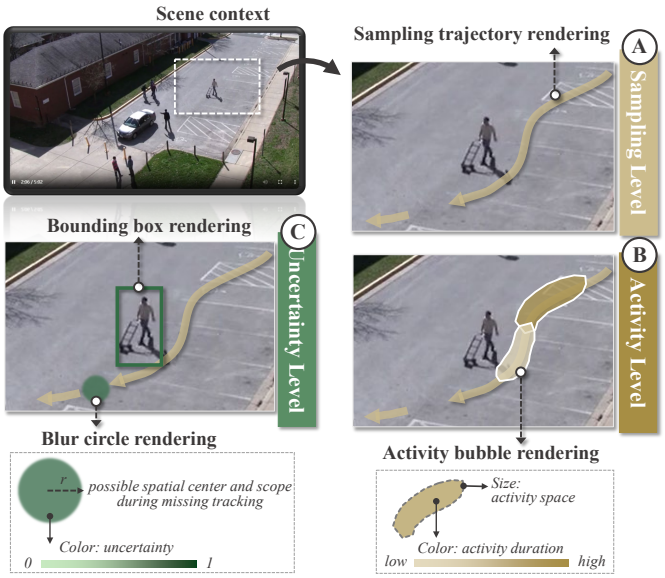


Fig. 8. On-demand rendering design for scene view. Trajectory lines (sampling level), activity bubbles (activity level), blur circles and bounding boxes (uncertainty level).

VI. Case Study

We present three use cases to evaluate LANDER. Three experts operated LANDER to explore pedestrian activities in various videos. We collected their direct feedback and evaluations, aiming to assess LANDER's effectiveness based on their expertise and experience.

A. Method

1) Experts: Three experts (E4, E5 and E6) were recruited to the case study. E4 (male, $age = 28$, $work = 5$), a junior security personnel, specializes in video surveillance. He has been responsible for reviewing pedestrian activities during multiple security incidents. His experience and workflow closely align with the demands of evaluating LANDER performance. E5 (female, $age = 28$), a PhD at our institute, specializes in the intersection of visualization and surveillance video research. She exhibits exceptional insight in understanding visualization technologies and their application in surveillance video analysis. E5 can offer valuable insights and analyses for our method. E6 (male, $age = 34$, $work = 6$), an associate professor at our institute, specializes in video vision analysis. He has a deep understanding of intelligent video technologies and domain issues, and can provide valuable advice for LANDER from a vision analysis perspective. Their involvement enriched our evaluation, providing insights into LANDER's technology, application, and user experience.

2) Preparation: Expert E4, unable to attend in person, conducted his case analysis online. We provided him with an external link to LANDER and recorded the procedure via Tencent Meeting. E4 and E5 conducted their analysis offline. Before starting, we allocated 15 minutes for them to familiarize themselves with LANDER. Then, they freely explored within LANDER based on assigned tasks. A 23-inch, 1920×1080 resolution monitor was used.

B. Case One: Individual activity in the parking lot

The video is VIRAT_S_000002.mp4 [54], was taken in a parking lot. The main exploration tasks of E4 were as follows: 1) identify *Tommy*'s activity time and area; 2) determine the category of his uncertainty. We summarized the main procedure of E4's exploration:

E4 initially explored the overview view (Fig. 9A). E4 noticed *Tommy*'s photo was the only one, while the man wearing a white hat appeared three times (**T1**). The differences triggered his thoughts. He pointed out, *"Tommy appears only once and his pixel bar occupies the entire timeline. This suggests that Tommy remains consistently present in the video and is not mistaken for someone else."* He further added, *"In other words, Tommy's activity data does not include **Category A**, and his information is complete."* (**T3**)

Later, E4 carefully observed the pixel bar chart. He found that the pixel bars in lower chart exhibited a trend of increasing height along the time axis (Fig. 9B): *"I speculate that Tommy is gradually approaching the main activity area in the video and may engage in social interactions with other pedestrians."* (**T5**). To verify this hypothesis, E4 randomly reviewed the scene view and operated the sampling trajectory rendering button and . He observed *Tommy* crossing the parking lot, briefly pausing by the building, and ultimately staying near the lawn (Fig. 9C). E4 summarized, *"My hypothesis is confirmed as Tommy consistently present. And in the later stage, his activity primarily focus on periphery of the parking lot, which is also a gathering place for pedestrians."* (**T6**)

In addition to the height changes in lower chart, E4 also mentioned the color variations in both upper and lower chart that caught his attention. He observed that the pixel colors in earlier phase are darker and exhibited significant fluctuations, while the colors in later phase are lighter and more stable (**T5**). He manipulated the uncertainty rendering button to analyze possible causes . E4 noticed that in the semantic timeline under video player, green color blocks are concentrated when *Tommy* is under the building, accompanied by unstable bounding box rendering (namely, intermittent rendering). In the spatial dimension, blur circles appear near the building (Fig. 9D). Conversely, during *Tommy*'s stay near the lawn, there are no green color blocks on the timeline, and the bounding box rendering remains consistent. No blur circles are observed in the area (Fig. 9E). E4 expressed excitement, *"In my view, it is very dark near the building, and even security personnel would find it difficult to discern pedestrian activity through video. The occlusion also makes it difficult for algorithms to recognize Tommy, resulting in **Category B** in his activity data. This also explains the emergence of visual elements symbolizing uncertainty."* (**T2, T3, T6**)

Subsequently, he conducted interactive visual analysis of *Tommy*'s action status. *Detailed content is elaborated in supplementary material Section V.*

C. Case Two: Item handover activity in the parking lot

The video is VIRAT_S_000101.mp4 [54], was taken in a parking lot. E5's main exploration task was to identify key time and location of item handover activity between *Allen* and *Jack*. We summarized the main procedure of E5's exploration:

Initially, E5 explored the overview view (Fig. 10A), noticing both *Allen* and *Jack*'s images appeared three times. She clicked to examine and discovered they frequently entered and exited the surveillance. She noted, *"When they re-entered the surveillance, the tracking algorithm failed to recognize them, introducing **Category A**."* Then, E5 clicked button to merge three segments of *Allen* and *Jack*'s activity (Fig. 10B), gaining a complete view of their activities. She observed overlapping activity times between them in all three segments. She clicked the button and to render the trajectory lines and activity bubbles (Fig. 10C). She discovered that they actually performed three item handovers near a car (Fig. 10D-E). E5 added, *"Relying solely on automatic assessment or manual video retrieval, I might have assumed the handover occurred only at the beginning, overlooking the latter two. In contrast, LANDER effectively alerted me to these occurrences."*

During exploration, E5 noticed an issue: despite frequent item handovers between the two, the "carry" bar is low (Fig. 10F). This indicates the action algorithm did not continuously recognition this activity. E5 commented, *This could be due to occlusion, rapid action occurrence, or incomplete training dataset."* E5 believed that focusing only on automatic assessment data might cause her to miss brief "carry" action records, potentially overlooking the item handover. In contrast, LANDER's advantage lies in its statistical analysis and spatio-temporal visualization of action information. This enabled E5 to quickly spot actions that were subtle in data, and easily locate them in the video for inspection.

D. Case Three: Group Activity Exploration in a street video

The video is fight_margaret_2.mp4 [55], was taken in a street. E6's main exploration task was to identify key group activities in the video. We summarized the main procedure of E6's exploration:

Expert E6 also showed great interest in the overview view (Fig. 11A). He noticed that three pedestrians' photos appeared repeatedly, while the other four appeared only once. E6 hypothesized: *"According to LANDER's design, this might indicate that these three pedestrians are involved in **Category A**."* E6 then reviewed the video segments and found that they indeed frequently entered and exited the surveillance. E6 noted: *"The algorithm failed to recognize their feature information, resulting in a failure to re-identify them."* E6 merged their data, obtaining a complete view of their activities and visual clues (Fig. 11B). He also noted that three pedestrians had larger pixel bar representations, indicating longer activities. In contrast, the other four pedestrians had shorter pixel bars, suggesting brief activities (Fig. 11E). Upon

Fig. 9. Case study One: E4's interaction exploration process for *Tommy*'s trajectory and related uncertainty.

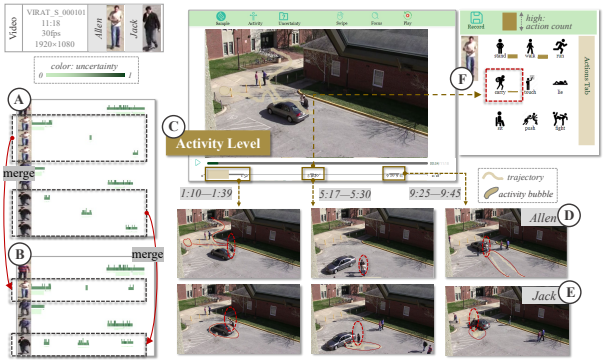


Fig. 10. Case study Two: E5's interaction exploration process for item handover activity between *Allen* and *Jack*.

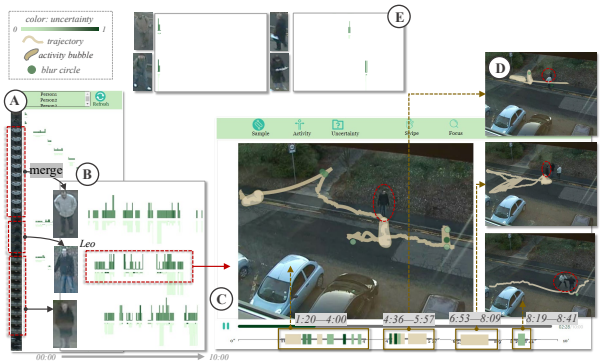


Fig. 11. Case study Three: E6's interaction exploration process for the group activity (walking and fighting) in the street video.

reviewing the footage, E6 discovered that the video mainly captured these three individuals' group activities, such as chatting or fighting (Fig. 11D).

Previously, E6 had activated all three level rendering buttons , and . The frequent color changes in *LEO*'s pixel bar and green blocks on the timeline drew E6's attention. This led E6 to notice frequent missing in *LEO*'s trajectory data, indicating **Category B**. E6 added, *"From my observation, LEO's attire isn't conspicuous, and he's often obscured by others or moves quickly, leading to unstable tracking."* E6 ultimately emphasized, based on his experience in computer vision, that vision algorithms struggle to achieve ideal results in such complex scenarios. He advised against overly relying on automatic assessment data but stressed that *"LANDER is a valuable tool for analysis support."*

E. Feedback

After case studies with experts, we received positive feedback and valuable suggestions for improving LANDER.

When asking for their views on the LANDER design: E4 particularly appreciated the scene rendering view. He emphasized, *"Connecting pedestrian activities and related uncertainties to the video scenes significantly improved analysis efficiency."* E5 found the exploration design from an overview to details is user-friendly. It allows for an overview of complex activity data while facilitating in-depth. However, she suggested that scene rendering view should support display of multi-person activity data. *"The current view is not ideal when it comes to comparative or joint analysis of multi-person activities,"* she pointed out. E6 found that LANDER's design incorporates some fundamental visualization elements suitable for most users. Particularly, he appreciated the use of icons instead of actions' textual descriptions, which was intuitive

and interesting. Furthermore, he noted that the dual-channel overview view might initially seem a bit complex.

When asked about integrating uncertainty visualization with pedestrian activity: E4 expressed appreciation and support. He believed that visualizing uncertainty linked with pedestrian activity aids in identifying when and where the algorithm underperforms. E5 found this approach innovative, noting that LANDER proposal fits with the concerns of video surveillance field. *"Few studies focus on uncertainty in automatic assessments, but LANDER introduces the analysis flow targeting common activity uncertainties."* E6 recognized the advantage of LANDER in visualizing uncertainties in assessment results. He believed it alerts users to potential errors. Additionally, it provides vision researchers with crucial clues about model performance, aiding in model or dataset optimization. He suggested incorporating uncertainty correction features in the analysis flow or modules. Such as action recognition result adjustments, allowing users to directly modify.

VII. User Evaluation

To further understand LANDER's effectiveness in facilitating exploration and insights into pedestrian activity and the uncertainty, we conducted a user evaluation. We followed three objectives to guide evaluation: **G1** evaluate LANDER's performance in pedestrian activity retrieval tasks. **G2** evaluate how well LANDER aids in understanding pedestrian activity and the uncertainty. **G3** evaluate user experience of LANDER.

A. Evaluation Method

1) Participants: We recruited 12 participants (3 females and 9 males, postgraduates, $age_{mean} = 24.5$, $age_{sd} = 3.2$) from fields including video surveillance, computer vision, and data visualization. The aim is to encompass various professional perspectives to evaluate LANDER's performance. P1-P3 (recommended by E1), with a solid background in video surveillance and expertise in video retrieval and investigation. We further invited 9 individuals in our affiliated research institution. P4-P7 specialize in video-related vision technology, with deep knowledge of video research, closely aligning with LANDER's workflow. P8-P12, focused to video visualization research, whose research achievements and experience can offer valuable insights into LANDER's design. These participants are appropriate for evaluating LANDER, as the user tasks primarily involve general retrieval rather than specific security investigation skills. Three participants with expertise in video surveillance can provide practical application insights, while others can offer technical and design perspectives.

2) Data and Apparatus: Videos involved are from the VIRAT [54] and BEHAVE [55] datasets. The videos have durations ranging from 10 to 60 minutes, rather than being mere summaries. They encompass diverse pedestrian activities in real-world surveillance scenarios, providing representative scenes for analysis. Due to reliance on vision models and lengthy pipeline, video data processing tasks were pre-executed on a Linux (Ubuntu 22.04.1) server. *For details on model deployment and virtual environment setup, please refer to the supplementary material Section III-A.* The server was equipped with an Intel Core i9-10900x CPU (3.70GHz) and a Nvidia GeForce RTX 2080 Ti GPU (11GB). In addition, LANDER was hosted on a desktop machine running Windows version 19045.3693, with an Intel Core i5-9400 CPU (2.9GHz). The user interface, displayed on a 23-inch monitor with a 1920×1080 resolution, enables real-time interaction.

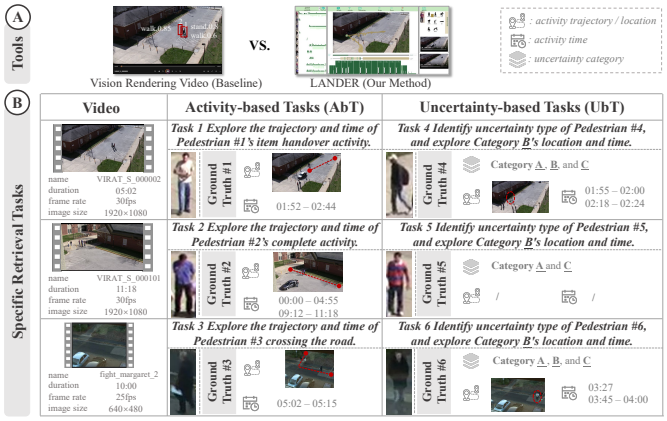


Fig. 12. The comparative evaluation based on Task1-Task6.

3) Comparative Evaluation: We first designed a user task-based comparative evaluation. It involves comparing LANDER with a baseline method (Fig. 12A), which is manually reviewing videos processed by vision algorithms, to evaluate differences in efficiency, accuracy and other factors (**G1**). Specifically, we required participants to perform six retrieval tasks using both LANDER and baseline method (Fig. 12B). The six retrieval tasks were categorized into two types: activity-based (AbT) and uncertainty-based (UbT). The AbT tasks focus on evaluating LANDER's performance in in activity localization and temporal analysis. Participants were asked to explore locations and time related to specific pedestrian activity. The UbT tasks emphasize evaluating LANDER's capability in recognizing and analyzing uncertainty. Participants were asked to identify uncertainty categories in pedestrian activity. Specifically, they need to pinpoint locations and times of **Category B**. For detailed information on video parameters, specific task designs, and ground truth, please refer to Fig. 12. To assess LANDER's performance in the retrieval tasks, we established four key metrics. These metrics quantitatively compare LANDER with the baseline method across temporal, spatial, and uncertainty dimensions (**G2**):

- *execution time*: Measures the total time taken by participants to complete a retrieval task, evaluating the efficiency in retrieval performance.
- *time error*: The discrepancy between participants' estimated activity time and the actual time, reflecting the ability to accurately identify the timing of activity.
- *trajectory error*: The difference between participants' described trajectories and the actual trajectories, assessing the capability in processing spatial information.
- *uncertainty category error*: The comparison between participants' identified uncertainty categories and the ground truth, measuring the performance in recognizing and understanding uncertainty.

4) User Questionnaire: To assess LANDER's efficacy in aiding user understanding of pedestrian activities and their uncertainties, as well as user experience (**G2**&**G3**), we also designed the questionnaire. The questionnaire comprised 17 questions (Fig. 14), focusing on LANDER's visualization, operation, and user experience:

- *Q1–Q3*: Gather user experiences with the tool, focusing on pleasure, confidence, and workload in operation.
- *Q4–Q6*: Gather evaluations on LANDER's effectiveness in analyzing pedestrian activities and uncertainties.
- *Q7–Q13*: Gather comments about the functionality and practicality of LANDER's visual designs.
- *Q14–Q17*: Collect user feedback on operational simplicity, learning curve, and interaction experience.

The questionnaire results were quantitatively scored using 7-ponit Likert Scale. It has gained widespread recognition in psychometrics, offering sufficient sensitivity to distinguish between attitudinal differences.

B. Procedure

Firstly, participants were introduced to the baseline videos, which displayed elements rendered by visual algorithms, including tracking boxes, ID labels, and confidence labels. After the introduction, they performed six tasks using the baseline method and completed questionnaire *Q1–Q3*.

Secondly, to mitigate potential biases in evaluating LANDER method following the baseline due to memory effects, we considered two programs: cross-evaluation and a cooling-off period. Given the limited sample size and the involvement of video variables, we opted for a program combining additional tasks with a cooling-off period. Hence, following the baseline method, participants were asked to execute additional tasks (*see supplementary material Section VI for details*). The results were not included in the final evaluation. Subsequently, the evaluation process entered a 24-hour cooling-off period [56].

Thirdly, we introduced participants to LANDER and provided a demonstration of how to use it. Similarly, after the instruction, participants used LANDER to perform six tasks and complete questionnaire *Q1–Q3*. Finally, participants were required to complete questionnaire *Q4–Q17*. All participants conducted our study face-to-face. Afterward, each participant was compensated with US $8.

C. Results

We first conducted statistical analyses on the results of six tasks and questionnaire *Q1–Q3*. We used error bar charts to compare the performance differences between LANDER and baseline method, as shown in Fig. 13. Given our sample size ($N = 12$), the non-normal distribution of data, and the comparison of baseline and our method under consistent tasks and participants, we chose the Wilcoxon signed-rank test. It is a widely used non-parametric method, suitable for small samples and paired sample comparisons. The G-test is suitable for categorical data with non-normal distributions, particularly exhibiting robustness in small samples or situations with zero frequencies. It can effectively assess user responses across different rating levels. The overall superiority of LANDER in metrics is shown in Table. I. Furthermore, we used bar charts to represent the questionnaire *Q4–Q17* results (Fig. 14), showcasing user experiences.

TABLE I: LANDER's performance relative to baseline (↑ increase, ↓ decrease), with p-values indicate significance.

Metrics	Ratio (%)	p-values
execution time	↓ 22.91%	$p < 0.01$
time error	↓ 22.27%	$p = 0.41$
trajectory error	↓ 48.15%	$p < 0.01$
uncertainty category error	↓ 68.25%	$p < 0.05$
Q1 (pleasure)	↑ 58.14%	$p < 0.01$
Q2 (workload)	↓ 38.71%	$p < 0.01$
Q3 (confidence)	↑ 18.27%	$p < 0.05$

In summary, for *execution time*, LANDER demonstrates a significant 22.91% efficiency increase compared to baseline ($p < 0.01$). However, the 22.27% reduction in *time error* is not statistically significant ($p = 0.41$). For *trajectory error*, LANDER shows a 48.15% reduction, a statistically significant enhancement ($p < 0.01$). Moreover, a 68.25% reduction in *uncertainty category error* validates LANDER's effectiveness in uncertainty identification ($p < 0.05$). In user questionnaire, LANDER shows a 58.14% increase in user pleasure *Q1* ($p < 0.01$), a 38.71% decrease in workload *Q2* ($p < 0.01$), and an 18.27% increase in task confidence *Q3* ($p < 0.05$). The statistical results are analyzed in detail as follows:

As shown in Fig. 13A, LANDER notably reduces the execution time. Specifically, in AbT Task1-Task3, the average time is decreased by 22.93%; in UbT Task4-Task6, it is reduced by 22.89%. Participants noted that LANDER's visual presentation and analysis significantly improved retrieval efficiency.

As shown in Fig. 13B, LANDER does not demonstrate a significant advantage in *time error* metric. This is mainly due to task nature and data characteristics. In most tasks, activity times are clearly defined, making manual retrieval relatively effective. Although LANDER offers higher efficiency, its improvements in time localization accuracy are not pronounced. But in Task2, the error line in baseline method shows greater variability (*mean* = 22.17, *std* = 46.44). This is due to two participants mistakenly thought Pedestrian #2's activity ended after initially exiting the video. In fact, Pedestrian #2 reappeared towards the end of video, leading to significant errors. In such scenarios, manual retrieval is prone to errors due to missing information, whereas LANDER effectively avoids this issue by providing a global view.

As shown in Fig. 13C, LANDER's error in trajectory identification is about 48.15% lower than baseline. User feedback particularly highlights LANDER's advantages in presenting and analyzing pedestrian trajectories. Particularly in handling complex trajectories, it effectively prevents information loss. P1 stated:*"With the baseline method, I often forget the pedestrian's path, but LANDER effectively alleviates this problem."*

As shown in Fig. 13D, in uncertainty identification tasks, using LANDER reduces error by 68.25% compared to the baseline method. Most participants report that LANDER simplifies

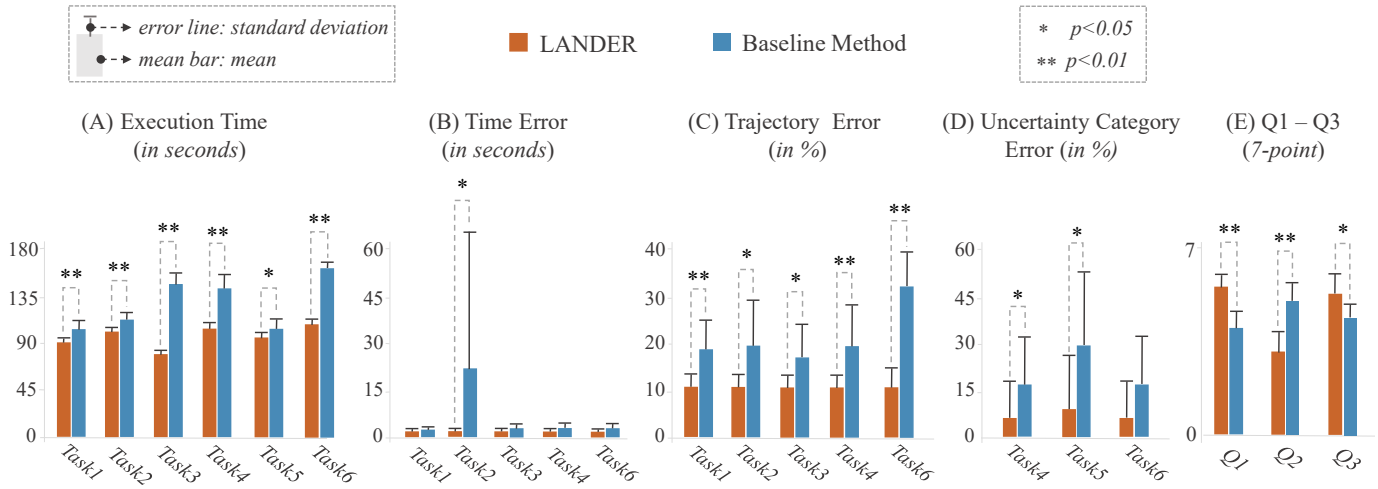


Fig. 13. Error bar charts for statistical results. Orange represents LANDER, while blue represents baseline method. In addition, *($p < 0.05$) and **($p < 0.01$) identify the statistical significance. No markers indicate a p-value beyond 0.05.

uncertainty detection, whereas the baseline method struggles in this aspect. P7 stated:*"The baseline method presents all pedestrian information at once, making it hard for me to focus on identifying uncertainties in activities."*

As shown in Fig. 13E, LANDER significantly improves user satisfaction compared to baseline method. P8 commented, *"I feel much more relaxed using LANDER."*. While P2 noted, *"I often resist the vision algorithm due to excessive uncertainty feedback, but LANDER makes my attitude more positive."*

As shown in Fig. 14, participants generally agree that LANDER's visual design effectively supports pedestrian activity and uncertainty analysis (*Q4–Q12*), especially its overview (*Q7–Q8*) and rendered views (*Q9–Q10*). They also expressed a strong desire to use LANDER in their future work (*Q6*). Participants highlight that LANDER's advantages are its *"intuitiveness"* and *"efficiency"*. It provides visual cues on spatio-temporal activity and uncertainty without involving the details of vision technology. 16.6% participants indicate that LANDER is insufficient to effectively support subsequent collaborative communication among security personnel (*Q13*).

Participants generally express satisfaction with LANDER's operating mechanism. 83% participants believe that learning cost of LANDER is low, and they can quickly get started (*Q15*). 25% participants (including P2, P5, and P6) note that understanding the multidimensional information encoded in dual-channel pixel chart takes time. All participants unanimously agree that LANDER is easy to operate, responsive, and offers a good user experience. P11 stated:*"clear interface layout and easily understandable operation"*.

VIII. Discussion and Future Work

Significance. Traditional video retrieval modes are often inefficient, especially when it comes to complex pedestrian activities. We introduce visualization technology into the surveillance video analysis process. This transforms complex activity data into an intuitive visual form, enabling a more streamlined and efficient analysis. However, many studies have neglected the impact of uncertainty in automatic vision-based surveillance video visualization. We specifically focus

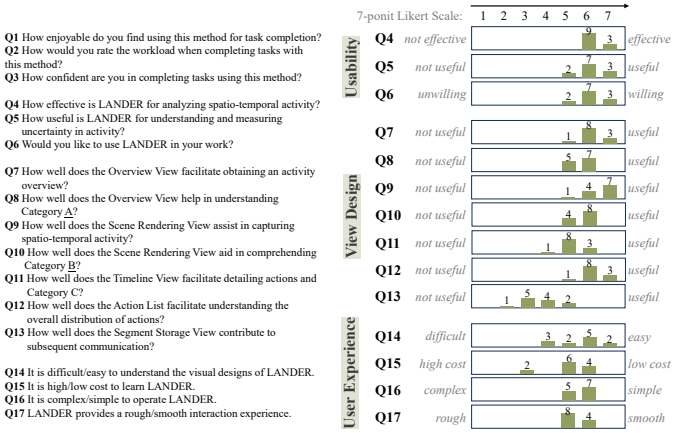


Fig. 14. Questionnaire content and bar charts of statistical results. Bar height indicates participant numbers.

on three uncertainties in automated activity assessment and explore how to visualize them to alert security personnel to potential risks. Domain experts also highly appreciate this idea, and feedback from user study demonstrates potential in communicating this information via visualization. Especially in the era of AI-generated content (AIGC), it's crucial to closely monitor the uncertainties introduced by AI methods in practical applications. This attention is particularly important in domains such as security, law, and healthcare.

Lessons learned. In collaboration with security personnel, we gained valuable insights from their feedback. Visual designers must customize designs to meet the specific needs of security personnel, considering their unique job requirements. We need to understand their workflow, data processing, and decision-making to provide visualization tools that align with their needs. Another important lesson learned is the importance of visual analysis. We realize that emphasizing visual analysis skills becomes particularly important when dealing with large volumes of video data. We need to ensure our interface provides *"clear overview and flexible functionality, without losing touch with the real context"*. Further, experts emphasized that *"the performance and response speed of the system are cru-*

cial". We need to ensure that LANDER's interactive real-time performance meets needs in practical scenarios. Measures such as asynchronous data processing, sliding window algorithm optimization for views, and efficient visualization rendering techniques can be considered to speed up the data rendering process on the interface. Additionally, the efficiency of video processing module can be improved through distributed model deployment and computing, lightweight modeling techniques, and GPU acceleration.

limitations and future work. There are limitations in our study that need to be discussed. Firstly, the scene rendering view in LANDER has not yet incorporated dynamic embedded visualization. To achieve dynamic layouts, we plan to develop a dynamic programming-based layout optimization algorithm. It is guided by principles of alignment, balance, and minimal obstruction to optimally place visual elements in video scenes. To enhance efficiency, we propose integrating a keyframe sampling strategy into existing scene rendering workflow. Afterward, we will apply the layout algorithm and leverage technologies like the GreenSock Animation and OpenGL to achieve dynamic rendering effects. Secondly, our study focuses on three categories of uncertainty. However, information omissions due to missing actions detection also represent a significant source of uncertainty. To response this, anomaly detection modules can be integrated to assess abnormalities in video frames. Combining these assessments with action recognition results quantifies uncertainty. Additionally, introducing user feedback mechanism allows analysts to annotate missed actions, facilitating interactive supplementation and updates. Furthermore, we were unable to access surveillance video data from law enforcement agencies due to the sensitive nature of data. This limited our understanding of LANDER's robustness. Future research could consider more extensive collaborations to obtain diverse and authentic scenario data, enabling a more comprehensive evaluation of LANDER's performance.

IX. Conclusion

This study explores the visualization solution for pedestrian activity and uncertainty analysis. Based on a formative study, we propose a design framework that integrates spatio-temporal activity data analysis, uncertainty modeling, and interactive visual exploration. We also develop the user interface LANDER to implement the concept of this design framework. Feedback from expert interview and user study demonstrates LANDER's strong performance, affirming its effectiveness and usability. In future work, we plan to improve and expand LANDER's visual design to enhance its data presentation and analysis capabilities. We will further evaluate LANDER's performance in real scenarios to guide its optimization and enhancement.

Acknowledgments

This work is supported by Zhejiang Provincial Natural Science Foundation of China (LR23F020003, LTGG23F020005), National Key Research and Development Program of China (2022YFB3104800), National Natural Science Foundation of China (62372411, 62036009), Fundamental Research Funds for the Provincial Universities of Zhejiang(RF-B2023006).

References

[1] Teddy Ko. A survey on behaviour analysis in video surveillance applications. *Video Surveillance*, 1(1):279–294, 2011.

[2] Erik Blasch, Zhonghai Wang, Haibin Ling, Kannappan Palaniappan, Genshe Chen, Dan Shen, Alex Aved, and Guna Seetharaman. Video-based activity analysis using the l1 tracker on virat data. In *Proceedings of the Applied Imagery Pattern Recognition*, pages 1–8, 2013.

[3] Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, 77(1):103–116, 2021.

[4] Rita Borgo, Min Chen, Ben Daubney, Edward Grundy, Gunther Heidemann, Benjamin Höferlin, Markus Höferlin, Heike Jänicke, Daniel Weiskopf, and Xianghua Xie. A survey on video-based graphics and video visualization. In *Proceedings of the European Association for Computer Graphics*, pages 1–23, 2011.

[5] Benjamin Höferlin, Markus Höferlin, Gunther Heidemann, and Daniel Weiskopf. Scalable video visual analytics. *Information Visualization*, 14(1):10–26, 2015.

[6] Markus Höferlin, Benjamin Höferlin, Daniel Weiskopf, and Gunther Heidemann. Uncertainty-aware video visual analytics of tracked moving objects. *Journal of Spatial Information Science*, 2011(2):87–117, 2011.

[7] Ralf P Botchen, Sven Bachthaler, Fabian Schick, Min Chen, Greg Mori, Daniel Weiskopf, and Thomas Ertl. Action-based multifield video visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):885–899, 2008.

[8] Markus Höferlin, Benjamin Höferlin, and Daniel Weiskopf. Video visual analytics of tracked moving objects. In *Proceedings of the Workshop on Behaviour Monitoring and Interpretation*, pages 59–64, 2009.

[9] Philip DeCamp, George Shaw, Rony Kubat, and Deb Roy. An immersive system for browsing and visualizing surveillance video. In *Proceedings of the ACM International Conference on Multimedia*, pages 371–380, 2010.

[10] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Proceedings of the European Conference on Computer Vision*, pages 850–865, 2016.

[11] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018.

[12] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019.

[13] Fengwei Gu, Jun Lu, and Chengtao Cai. Rpformer: A robust parallel transformer for visual tracking in complex scenes. *IEEE Transactions on Instrumentation and Measurement*, 71(1):1–14, 2022.

[14] Fengwei Gu, Jun Lu, Chengtao Cai, Qidan Zhu, and Zhaojie Ju. Repformer: A robust shared-encoder dual-pipeline transformer for visual tracking. *Neural Computing and Applications*, 35(28):20581–20603, 2023.

[15] Fengwei Gu, Jun Lu, Chengtao Cai, Qidan Zhu, and Zhaojie Ju. Eantrack: An efficient attention network for visual tracking. *IEEE Transactions on Automation Science and Engineering*, 1(1):1–18, 2023.

[16] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3645–3649, 2017.

[17] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *Proceedings of the IEEE International Conference on Image Processing*, pages 3464–3468, 2016.

[18] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023.

[19] Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. Motiontrack: Learning robust short-term and long-term motions for multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17939–17948, 2023.

[20] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 568–576, 2014.

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional

networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[22] Junwu Weng, Chaoqun Weng, and Junsong Yuan. Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4171–4180, 2017.

[23] Anshul Shah, Aniket Roy, Ketul Shah, Shlok Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18846–18856, 2023.

[24] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18011–18021, 2023.

[25] Zan Gao, Xinglei Cui, Tao Zhuo, Zhiyong Cheng, An-An Liu, Meng Wang, and Shenyong Chen. A multitemporal scale and spatial-temporal transformer network for temporal action localization. *IEEE Transactions on Human-Machine Systems*, 53(3):569–580, 2023.

[26] Chunggi Lee, Yeonjun Kim, Seungmin Jin, Dongmin Kim, Ross Maciejewski, David Ebert, and Sungahn Ko. A visual analytics system for exploring, monitoring, and forecasting road traffic congestion. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3133–3146, 2019.

[27] Gromit Yeuk-Yin Chan, Luis Gustavo Nonato, Alice Chu, Preeti Raghavan, Viswanath Aluru, and Cláudio T Silva. Motion browser: Visualizing and understanding complex upper limb movement under obstetrical brachial plexus injuries. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):981–990, 2019.

[28] Linmi Tao and Meiqing Zhang. Understanding an online classroom system: Design and implementation based on a model blending pedagogy and hci. *IEEE Transactions on Human-Machine Systems*, 43(5):465–478, 2013.

[29] José A Ruipérez-Valiente, Pedro J Munoz-Merino, José A Gascón-Pinedo, and C Delgado Kloos. Scaling to massiveness with analyse: A learning analytics tool for open edx. *IEEE Transactions on Human-Machine Systems*, 47(6):909–914, 2016.

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

[31] Haotian Zhang, Cristobal Sciutto, Maneesh Agrawala, and Kayvon Fatahalian. Vid2player: Controllable video sprites that behave and appear like professional tennis players. *ACM Transactions on Graphics*, 40(3):1–16, 2021.

[32] Zhutian Chen, Shuainan Ye, Xiangtong Chu, Haijun Xia, Hui Zhang, Huamin Qu, and Yingcai Wu. Augmenting sports videos with viscommentator. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):824–834, 2021.

[33] Jeongyeon Kim, Daeun Choi, Nicole Lee, Matt Beane, and Juho Kim. Surch: Enabling structural search and comparison for surgical videos. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[34] Kanglei Zhou, Ruizhi Cai, Yue Ma, Qingqing Tan, Xinning Wang, Jianguo Li, Hubert PH Shum, Frederick WB Li, Song Jin, and Xiaohui Liang. A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2456–2466, 2023.

[35] Zhutian Chen, Qisen Yang, Jiarui Shan, Tica Lin, Johanna Beyer, Haijun Xia, and Hanspeter Pfister. iball: Augmenting basketball videos with gaze-moderated embedded visualizations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[36] Yuchen Wu, Yuansong Xu, Shenghan Gao, Xingbo Wang, Wenkai Song, Zhiheng Nie, Xiaomeng Fan, and Quan Li. Liveretro: Visual analytics for strategic retrospect in livestream e-commerce. *arXiv preprint arXiv:2307.12213*, 2023.

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

[38] Weijia Xu, Natalia Ruiz, Kelly Pierce, Ruizhu Huang, Joel Meyer, and Jen Duthie. Detecting pedestrian crossing events in large video data from traffic monitoring cameras. In *Proceedings of the IEEE International Conference on Big Data*, pages 3824–3831, 2019.

[39] Shijing Han, Huadong Wang, Xiangyang Hao, and Shufeng Miao. 3d geographic trajectories' generation and visualization of dynamic objects in surveillance videos. In *Proceedings of the International Conference on Image, Vision and Computing*, pages 669–673, 2022.

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

[41] Mario Romero, Jay Summet, John Stasko, and Gregory Abowd. Viz-a-vis: Toward visualizing video through computer vision. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1261–1268, 2008.

[42] Mario Romero, Alice Vialard, John Peponis, John Stasko, and Gregory Abowd. Evaluating video visualizations of human behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1441–1450, 2011.

[43] Tamara Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[44] Haipeng Zeng, Xinhuan Shu, Yanbang Wang, Yong Wang, Liguo Zhang, Ting-Chuen Pong, and Huamin Qu. Emotioncues: Emotion-oriented visual summarization of classroom videos. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3168–3181, 2020.

[45] Haotian Li, Min Xu, Yong Wang, Huan Wei, and Huamin Qu. A visual analytics approach to facilitate the proctoring of online exams. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[46] Min Chen, Ralf Botchen, Rudy Hashim, Daniel Weiskopf, Thomas Ertl, and Ian Thornton. Visual signatures in video visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1093–1100, 2006.

[47] Ralf P Botchen, Min Chen, Daniel Weiskopf, and Thomas Ertl. Gpu-assisted multi-field video volume visualization. In *Proceedings of the Volume Graphics*, pages 47–54, 2006.

[48] Carter de Leo and Bangalore S Manjunath. Multicamera video summarization and anomaly detection from activity motifs. *ACM Transactions on Sensor Networks*, 10(2):1–30, 2014.

[49] Chingtang Fan, Yuankai Wang, and Cairen Huang. Heterogeneous information fusion and visualization for a large-scale intelligent video surveillance system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):593–604, 2016.

[50] Guodao Sun, Tong Li, and Ronghua Liang. Survizor: visualizing and understanding the key content of surveillance videos. *Journal of Visualization*, 25(2):635–651, 2022.

[51] Giovani Valdrighi, Nivan Ferreira, and Jorge Poco. Morevis: A visual summary for spatiotemporal moving regions. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):1–13, 2023.

[52] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[53] Alan M MacEachren, Robert E Roth, James O'Brien, Bonan Li, Derek Swingley, and Mark Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505, 2012.

[54] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3153–3160, 2011.

[55] Scott Blunsden and RB Fisher. The behave video dataset: Ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(4):1–12, 2010.

[56] Hermann Ebbinghaus. Memory: A contribution to experimental psychology. *Annals of Neurosciences*, 20(4):155–157, 2013.