Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 2203–2213
https://doi.org/10.54941/ahfe1005788

Auto-Generating Road Trip Vlogs While Safe-Driving: A Human-Vehicle-Environment System for Capturing and Editing Scenic Views En Route

Zichun Guo[1], Xianning Meng[2], Haiqing Xu[3], Zumeng Liu[4], Lingkan Wang[5], Jiawei Chen[6], and Yaxin Zhu[7]

[1]Beijing University of Chemical Technology, China
[2]Daegu University, FL 32816, Korea
[3]Georgia Institute of Technology, USA
[4]Imperial College London, UK
[5]Carnegie Mellon University, USA
[6]Beijing Normal University, China
[7]Tsinghua University, China

ABSTRACT

Road trip vlogs have gradually become popular content for sharing among people. This study introduces an artificial intelligence (AI)-based road trips video editing system, designed with a primary purpose of preventing traffic accidents caused by drivers recording scenic views with smartphones while driving. To enable the dashcam to automatically capture materials of interest, it is crucial to define the starting-ending time and content. Data monitoring around the Human-Vehicle-Environment (HVE) is a critical factor for establishing the capture rules. Guided by audiovisual language theory, the captured video materials are used to support intelligent editing. Furthermore, stylization in editing, including narrative lyrical and documentary style, is another design factor to achieve diversity in videos. Research results demonstrate that the synergy among the HVE elements is a pivotal factor in capturing key visuals. Using AI to complete road trip videos can reduce traffic accident risks and promote the effectiveness of recording scenic views.

Keywords: Human vehicle interaction, Auto video editing, Road trip

INTRODUCTION

With social media's profound influence, society increasingly documents and broadcasts personal experiences, highlighting connections and digital persona construction. Responding to this trend and the HVE paradigm, industries have developed specialized camera systems for various contexts: DJI's drones for aerial cinematography, and Insta360 for VR content. As road trips gain prominence in shared content, the convergence of automotive technology and intelligent video generation has become a interesting area in HCI fields.

Two main methods capture scenic drives. The first uses dashboard cameras to record entire trips, resulting in hours of footage. This exhaustive method presents storage and post-processing challenges and is not ideal for quick sharing or short snippets. The second involves drivers sporadically using smartphones, posing safety hazards. Distracted driving, a major cause of vehicular accidents, results in about 1.3 million traffic-related deaths annually (WHO, 2018). Solo driving trips in China have highlighted these safety concerns. In July 2023, a Nanjing driver using a head-mounted camera failed to prevent an accident due to delayed reaction (Mi, 2020). Addressing the balance between scenic captures and driver safety is a pressing HCI issue, especially with in-vehicle intelligent imaging.

In this investigation, we explore the confluence of human, vehicle, and environment (HVE) to develop a system that autonomously captures, processes, and curates travel-centric VLOGs. Our research focuses on the nuanced interactions within the HVE triad, designing vehicles that reflect the driver's preferences. We propose an intelligent road trip video framework that resonates with the multidimensional dynamics of HVE. Our journey includes:
1) designing a trajectory that integrates rich data from various driving phases, including the driver's emotions, the vehicle's behavior, and the surrounding environment; 2) creating a design model that aligns the driver's physiological and emotional states with vehicular metrics and environmental cues, forming a comprehensive system for in-vehicle video synthesis. Our empirical setting is the scenic Beijing-to-Inner Mongolia corridor, known for its breathtaking vistas and elite self-driving experiences in China. Safety is paramount; thus, our initial design iterations were refined within controlled game simulators before real-world testing.

In our study, we explored video generation through the HVE triad. We developed an on-board system for travel vlogs and used sensors to analyze drivers' responses. Our goal is to create autonomous vlogs that capture and present unique travel experiences. By analyzing drivers' reactions to sceneries, we identified captivating views through the interplay of the driver, vehicle, and environment. Our aim is to craft travel vlogs that reflect the driver's personal tastes, capturing and polishing travel stories in harmony with their preferences. The main contributions of our research are: 1) Feature Definement: Adopting a user-centric approach, we identify a comprehensive feature framework for high-quality travel videos, setting a gold standard for visual storytelling excellence. 2) Design Simplification: Simplify the design of elastic flat materials with structures by providing components that generate filled curves. 3) HVE Integration: Our method integrates HVE dynamics with picturesque video captures, using sensor data and advanced computer vision to create premium video content that captures peak moments. 4) Editorial Ingenuity: We introduce a new design ethos for in-vehicle travel video editing based on audiovisual linguistics, offering three distinct narrative styles, each with a unique flavor.

RELATED WORK

Road Trip Vlogs

Short travel videos cater to travelers' needs for self-expression, social interaction, and emotional support, shaping travel intentions, destination image perceptions, and travel experience evaluations (Iis 2009, Iis 2020). Research examines content selection, filming techniques, editing methods, and sharing platforms for travel vlogs. Key studies include context- aware video recommendation systems using data mining (Zhang, 2017), automatic road trip summary videos from dashcam footage (Bito, 2021), and collaborative 360° travel videos using navigational technologies (Kartikaeya, 2022). Conversely, substantial discourse exists on the impact of travel vlogs on creators and audiences, exploring travel experiences, identity construction, emotional attitudes, travel intentions, and destination image perceptions. Studies have examined how China's youth use platforms like TikTok or Douyin to document and share travel experiences, revealing motivations, content creation processes, dissemination strategies, and the impact on travel narratives and identity (Du, 2022). These short videos serve as travel mementos and mediums, facilitating connections, self-expression, and image crafting.

Human-Vehicle Interaction

In the domain of human-vehicle interaction, academic insights complement efforts in autonomous technologies and in-vehicle intelligent systems. Proprietary methodologies often obscure widespread understanding of driver monitoring systems in industrial domains (Zheng, 2022). For instance, autonomous driving technologies leverage sensor-based systems since the 1980s to enhance vehicles' perception and decision-making in traffic environments (Yeong, 2021). Efforts also extend to improving race safety through video data from race vehicle cameras and automated image recognition methods (Akkas, 2019). Meanwhile, early concepts from 1913 envisioned mounting cameras for recording journeys or entertainment, highlighting ongoing challenges with generating engaging content from lengthy dashcam recordings (Bito, 2021). Enhancing audiovisual design in intelligent editing is pivotal to avoid mechanical video outputs devoid of storytelling and artistic value, emphasizing the importance of addressing these issues.

Smart Video Editing

In today's era of user-generated content, rapid video editing is increasingly crucial, with nearly all smartphones now equipped for editing capabilities (Chen, 2021). The evolution of video editing has progressed from physical to digital and non-linear methods, advancing to automatic keyframe extraction and beyond (Hua, 2004). Research in intelligent video editing focuses on automatic video summaries through techniques like keyframes (Zhu, 2020) and multi-modal networks (Xu, 2023). Video restoration techniques, employing deep learning, aim to enhance visual clarity and sharpness, utilizing convolutional neural networks for image restoration and event detection for retrieving coherent images from low-quality videos (Ding, 2022). Automatic video editing techniques rooted in deep learning produce high-quality edits (Huang, 2022), including stylized editing tailored to user preferences using models like LSTM-GAN. However, there remains a gap in techniques specifically tailored for self-driving scenarios, focusing on scene understanding and personalized editing outputs in that context.

INTELLIGENT AUTOMOTIVE TRAVEL EDITING SYSTEM

We've developed an intelligent system for editing road trip Vlogs, integrating hardware for HVE data acquisition and software for content refinement. Using diverse sensors, we capture drivers' physiological and emotional data, alongside vehicular dynamics and environmental factors (Figure 1). Informed by contemporary audiovisual theories, selected footage undergoes sophisticated editing to create compelling road trip videos.

Essential Features of High Quality Vlog

In this section, we use case studies and data analytics to establish criteria for high-quality travel videos. Metrics like view counts, likes, shares, and comments gauge a video's appeal, but often lack precision in identifying resonant moments. To address this, we leverage YouTube's Engagement Graph, which tracks viewer engagement throughout videos, helping us pinpoint compelling segments. Additionally, on platforms like Bilibili, we analyze audience interactions using metrics such as the "High Energy Progress Bar," which measures comment density and engagement levels. From evaluating 26 distinct video segments, we identify 16 salient visual features such as Streetscape, Mountain, Cloud, and Rocks. Notably, Sky (96%), Streetscape (65%), and Mountain (65%) emerge as predominant visual elements. Compositionally, techniques like Panorama (81%) and Leading Lines (50%) are highlighted as effective tools for enhancing video quality and viewer engagement. These insights provide actionable guidelines for creators aiming to capture captivating scenic footage in their travel videos.

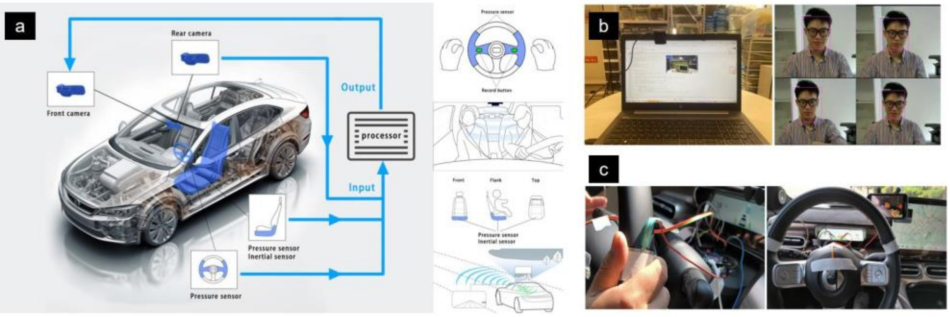


Figure 1: a) System overview of the HVE systerm b) System overview of the hardware components. c) Side view showing the computer with a camera setup for user monitoring.

HVE Collaborative Detection System

To achieve autonomous recognition and documentation of landscapes during transit, we integrate the "Human-Vehicle-Environment" (HVE) triad philosophy (Joshua, 2012). This approach emphasizes the interactions among humans, vehicles, and the environment, spanning disciplines like traffic engineering, computer science, and human-computer interaction. Studies highlight the importance of interactive methodologies in understanding users' relationships with physical and societal contexts (Paul, 2004). And research has focused on deploying sensor networks to evaluate these interactions, aiming to assess the interactions between humans, media, and their surroundings (Intille, 2006). Regarding the collaborative design of HVE, investigations can be partitioned across three distinct dimensions:

(i) Human: Delving into physiological, psychological, behavioral, and emotional characteristics of drivers and passengers, along with their interaction modalities and outcomes with both vehicles and the environment.

(ii) Vehicle: Probing into the dynamics, control systems, constraints, and other attributes of the vehicle, as well as their reciprocal actions and implications with humans and the environment. (iii) Environment: Assessing factors like roads, natural settings, weather, and other extrinsic variables that impact human-vehicle interactions.

Hardware Design

Based on the human-vehicle-environment framework mentioned above, we have developed a preliminary sensing system for capturing user, vehicle, and environmental information during the driving process. This detection system is designed to assist in capturing frames of interest to users during their travel experiences.

Human. Many papers related to human-vehicle interaction and autonomous driving have explored systems to track users' physiological, psychological, and behavioral data for smart driving. In this article, we primarily focus on user's the psychological impact and potential behaviors of the surrounding scenery during the road trip. We deploy multiple cameras to capture facial expressions and upper body posture for generating stylized video clips when significant changes are detected. However, recognizing that users typically show minimal facial expressions while driving alone, we also utilize physiological and behavioral data from sensors to infer their emotional state. Pressure sensors on the steering wheel and seats detect grip changes, indicating whether the user is relaxed or tense. Low pressure suggests relaxation, while high pressure indicates tension. This data, alongside other sensor readings, helps identify moments of significant scenery changes. Body posture, monitored by sensors on the back and seat, further informs the user's psychological state. Additionally, wearable physiological monitors track heart rate, respiration, and other signals automatically to assess the user's emotional state during driving.

Vehicle. When assessing the vehicle's condition, we consider its operational state and in-car equipment status, including the number, activation, angles, and positions of onboard cameras. This helps determine the system's capability to capture video clips effectively. For instance, evaluating the vehicle’s stability via acceleration guides decisions on when to capture frames, especially on rough roads prone to camera shake. Moreover, the vehicle’s operational data indirectly reflects the driver’s condition. Speed and steering responses vary with road conditions and driver alertness, influencing the video style—whether capturing fleeting urban scenes at high speeds or serene landscapes on smoother routes. Using a gyroscope on the steering wheel, we monitor steering movements and angular velocity to gauge the vehicle’s operational state during filming.

Environment. Environmental detection in car travel primarily focuses on capturing external scenery, essential for video travel recording. Dashcams are typically used to record images ahead and behind the vehicle. Factors such as camera placement, resolution, and field of view significantly influence automated video editing. Multiple cameras can capture landscapes from various angles, including views on both sides of the vehicle, enhancing the diversity of footage. Capturing the interior view of the vehicle is symbolic in Vlogs, reflecting the driver’s actions and expressions during the journey. In this experiment, we concentrated on editing footage from the front-facing camera. Computer vision algorithms facilitate the analysis, identification, and categorization of these images. Drawing from the analysis of key visual elements in exemplary travel videos, detailed rules for extracting environmental content will be outlined in subsequent sections.

Automated Editing System

Our automated editing system is developed in Python 3.11.1, viewing the smart environment as a discrete system. Data security is prioritized with storage within individual vehicles, accessible from external sources. Leveraging audiovisual language theory, our system adeptly processes HVE data to create compelling travel vlogs, ensuring the journey’s essence is engagingly presented to viewers.

Environment Classification. The essence of a video lies in its ability to narrate effectively, categorized broadly into process statements and stasis statements [57] . Using semantic segmentation models like UNet, we perform pixel-level semantic predictions to classify video materials into these categories. Process statements emphasize documentary-style scenes such as street views and cruising boats, while stasis statements feature emotive visuals like vast grasslands, evening glow, and sunsets. This approach allows us to craft a balanced and evocative video composition that resonates with viewers.

Feature Extraction of Audio-Visual Language. We streamline video materials using audio-visual features. Our scene classification module combines ResNet for feature extraction with BiLSTM to model video sequences, discerning between ’environmental’, ’narrative’, and ’detail’ scenes based on human presence. Scene labels generate descriptive sequences. Meanwhile, a convolutional neural network in the composition module evaluates visual balance, symmetry, and adherence to the rule of thirds in travel photos, refining aesthetic rules through extensive datasets and image augmentation techniques.

Editing Paradigm. Our editing module comprises two key components: the narrative editing module and the expressive editing module. The narrative editing module focuses on assembling event video clips to construct a coherent narrative based on script logic, prioritizing material from process statements over stasis statements. Conversely, the expressive editing module emphasizes predicting emotional and aesthetic features of video content, allowing users to specify desired emotional or stylistic attributes. Here, stasis statements play a more significant role compared to process statements. These modules enable the creation of three distinct video styles based on the balance between process and stasis statements: 1) Narrative Style: This style integrates predominantly narrative scenes from the process narrative repository, arranged in a logical timeline. Environmental and detail scenes complement the narrative flow. 2) Lyrical Style: In this style, fewer narrative scenes are used from the process narrative, while there is a greater emphasis on environmental and detail scenes from the stasis narrative. The editing focuses on creating a rhythmic, serene composition with seamless transitions.
3) Documentary Style: This raw style directly incorporates clips from the process narrative archive, presenting an unfiltered and authentic experience. It aims to immerse viewers in the journey without extensive narrative or expressive modifications.

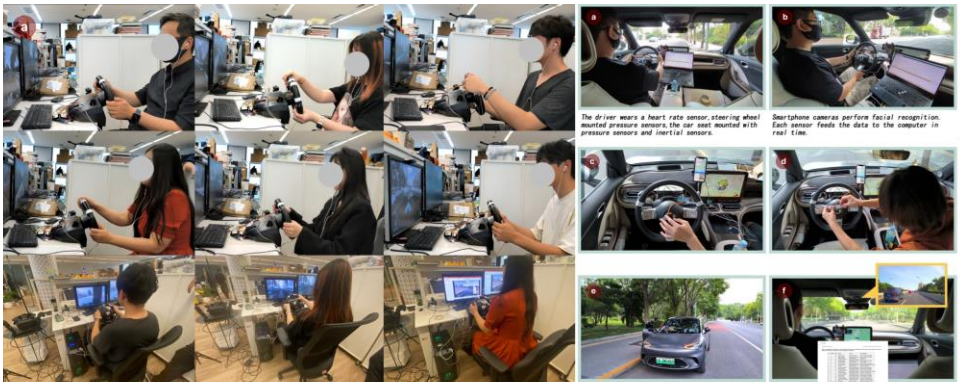
Evaluation

In this study, our focus is on detecting users' physiological, behavioral, and emotional responses to significant scenes and understanding their experience with our sensing system.

Participants. We recruited 10 participants aged 19 to 35 (6 female, 4 male) from the general population, all with driving experience. Seven participants have experience in travel videography, while one participant has experience with 3D motion sickness.

Experiment Setup. We chose "Horizon 4" as our 3D simulation platform
to authentically emulate driving experiences. Our setup included a driving simulator with pressure sensors on the steering wheel and angular velocity sensors for steering input. Two cameras were used: Camera A captured the participant's face, upper body, and in-game screen from the left, while Camera B focused on facial expressions from the front. Gameplay footage was recorded continuously. Experimenter A interacted with participants, and Experimenter B documented instances of interest and relevant data. (Figure 3).

Procedure. Participants provide detailed feedback on intriguing events, preferred camera angles, and recording durations. The experiment begins with baseline data collection and a briefing on the route and scenic highlights. After a brief orientation to the simulation mechanics, participants drive for 10 to 15 minutes while data from two cameras, a pressure sensor, and an angular velocity sensor are logged with timestamps. Our sensor suite captures physiological data (e.g., heart rate), behavioral data (e.g., steering wheel pressure), facial expressions, and vehicular telemetry to gain comprehensive insights. Post-drive, semi-structured interviews gather feedback on the sensing system and suggestions for improving automated road trip editing. The entire experiment lasts about 25 to 35 minutes, including driving and feedback sessions.



The driver wears a heart rate sensor, steering wheel mounted pressure sensors, the car seat mounted with pressure sensors and inertial sensors.

Smartphone cameras perform facial recognition. Each sensor feeds the data to the computer in real time.

Figure 2: User study - Participant experiencing our system. The study examines physiological, behavioral, and emotional responses to significant scenes and evaluates user interactions with the sensing system.

RESULTS

Preference of Scence Element During the Road Trip. We gathered data from 10 participants who specified moments they desired to capture during driving, totaling 74 reported instances. Analyzing game footage, we categorized these moments into buildings (20 instances), skies (15 instances), forests and lakes (12 instances), mountains (7 instances), rain and snow (7 instances), objects (5 instances), hybrid landscapes (8 instances), and scenes with significant changes (10 instances). These categories mirror findings from our earlier high-quality Vlog analysis, indicating user visual preferences during real driving experiences. Moreover, 80% of participants reported capturing scenes with buildings, while 70% mentioned aspects like sky and terrain. Additionally, 50% expressed comfort in scenes with significant changes, such as transitioning from forests to open areas. Unexpected elements like hot air balloons and fireworks were noted by P6 as memorable and worth capturing. Interestingly, only 2 participants focused on scenes on both sides of the car windows, with most concentrating on scenes ahead of them.

Analysis of the Log Data During the Driving Simluation. We documented user preferences for specific video segments and correlated these with sensor data from both users and the vehicle to explore predictive patterns. Key metrics included hand pressure on the steering wheel, vehicle speed, angular velocity, and user heart rate. Each participant's drive averaged 13 minutes and 31 seconds, with an average heart rate of 71 bpm and an average simulated speed of 80 km/h. We anticipated localized trends in user data during moments of interest. While heart rate analysis yielded no significant findings, hand pressure on the steering wheel showed an average value of 90 milliohms.

In 43% of instances, there was a noticeable peak (within 15 milliohms) around moments of interest, indicating tighter grip, possibly linked to heightened attention. Winding road segments also correlated with higher average pressure values, suggesting increased driver tension. Additionally, analyzing facial expressions from video footage revealed positive expressions 27 times, with 40% occurring at points of interest. These findings suggest a potential link between physiological responses and scenic engagement during simulated driving scenarios.

DISCUSSION

The High Visual Experience (HVE) system stands at the forefront of smart editing in vehicular environments, integrating advanced technology to enhance user experience. It addresses challenges in scenic capture and automated editing with a focus on transparency and adaptability in decision-making processes. User feedback underscores the demand for intuitive interfaces and user control, prompting inclusive design philosophies. In its approach to capturing and editing travel vlogs, the HVE system emphasizes narrative structure, aesthetic framing, emotional resonance, and authenticity to evoke viewer engagement. By integrating Human-Vehicle- Environment (HVE) interaction, the system enhances accuracy in content capture, considering physiological and emotional responses to scenic stimuli. The application of audiovisual language theory optimizes editing practices, catering to diverse styles and enhancing viewer immersion in travel narratives. Furthermore, the system accommodates cross-media elements such as aerial and mobile-captured content, facilitating comprehensive storytelling in travel filmmaking. As automobile designers advance in auto-editing applications for travel videos, integrating these elements within the HVE paradigm ensures a holistic approach to creating compelling and immersive travel experiences.

CONCLUSION

This study developed an intelligent onboard editing system for travel vlogs, enhancing scenic capture while ensuring driver safety. Using Human- Vehicle Environment (HVE) data and onboard sensors, we analyzed drivers' physiological and emotional responses to scenery, identifying captivating views. Applying audiovisual language theory, we crafted travel vlogs in three distinct styles, reflecting intelligent driving and societal contexts. An optimal onboard camera system should not only perceive environments but also understand driver emotions, road conditions, and landscapes to curate and edit compelling travel narratives. Achieving this requires continuous innovation from auto manufacturers, academia, and related sectors in the realm of intelligent technologies.

ACKNOWLEDGMENT

The Project Supported by National Social Science Foundation of China (Grant No. 23BZW014)

REFERENCES

De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. 2021. Sensor and sensor fusion technology in autonomous vehicles: A review. Sensors 21, 6 (2021), 2140.

Hsin-I Huang, Chi-Sheng Shih, and Zi-Lin Yang. 2022. Automated video editing based on learned styles using LSTM-GAN. In Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. 73–80.

Iis P Tussyadiah and Daniel R Fesenmaier. 2009. Mediating tourist experiences: Access to places via shared videos. Annals of tourism research 36, 1 (2009), 24–40. Iis Tussyadiah. 2020. A review of research into automation in tourism: Launching the Annals of Tourism Research Curated Collection on Artificial Intelligence and Robotics in Tourism. Annals of Tourism Research 81 (2020), 102883.

Jiaming Zhang, Yipeng Zhou, Di Wu, and Chunfeng Yang. 2017. Context-aware Video Recommendation by Mining Users' View Preferences Based on Access Points. In Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video. 37–42.

Joshua McVeigh-Schultz, Jennifer Stein, Jacob Boyle, Emily Duff, Jeff Watson, Avimaan Syam, Amanda Tasse, Simon Wiscombe, and Scott Fisher. 2012. Vehicular lifelogging: New contexts and methodologies for human-car interaction. In CHI'12 Extended Abstracts on Human Factors in Computing Systems. 221–230.

Kana Bito, Itiro Siio, Yoshio Ishiguro, and Kazuya Takeda. 2021. Automatic Generation of Road Trip Summary Video for Reminiscence and Entertainment using Dashcam Video. In 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. 181–190.

Kartikaeya KUMAR, Lev PORETSKI, Jiannan LI, and Anthony TANG. Tourgether360: Exploring 360 tour videos with others. (2022). CHI EA 22, 1–7.

Mifangjie. 2020. Driving down the highway enjoying the viewt. Retrieved July 3, 2023 from https://baijiahao.baidu.com/s?id=1677494741162586308&wfr=spider&for=pc

Paul Dourish. 2004. What we talk about when we talk about context. Personal and ubiquitous computing 8 (2004), 19–30.

Selahattin Akkas, Sahaj Singh Maini, and Judy Qiu. 2019. A fast video image detection using tensorflow mobile networks for racing cars. In 2019 IEEE International Conference on Big Data (Big Data). IEEE, 5667–5672.

Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2014. World-stabilized annotations and virtual scene navigation for remote collaboration. In Proceedings of the 27th annual ACM symposium on User interface software and technology. 449–459.

Stephen S Intille, Kent Larson, etc. 2006. Using a live-in laboratory for ubiquitous computing research. In Pervasive Computing: 4th International Conference, PERVASIVE 2006, Dublin, Ireland, May 7–10, 2006. Proceedings 4. Springer, 349–365.

Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. 2020. Dsnet: A flexible detect- to-summarize network for video summarization. IEEE Transactions on Image Processing 30 (2020), 948–962.

World Health Organization. 2018. road traffic injury. Retrieved July 2, 2023 from Website: https://www.who.int/zh/news-room/fact-sheets/detail/roadtraffic- injuries

Wujiang Xu, Runzhong Wang, etc. 2023. MHSCNET: A Multimodal Hierarchical Shot-Aware Convolutional Network for Video Summarization. In ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 1–5.

Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. 2004. Optimization-based automated home video editing system. IEEE Transactions on circuits and systems for video technology 14, 5 (2004), 572–583.

Xin Ding, Tsuyoshi Takatani, Zhongyuan Wang, Ying Fu, and Yinqiang Zheng. 2022. Event-guided Video Clip Generation from Blurry Images. In Proceedings of the 30th ACM International Conference on Multimedia. 2672–2680.

Xin Du, Toni Liechty, Carla A Santos, and Jeongeun Park. 2022. ‘I want to record and share my wonderful journey’: Chinese Millennials’ production and sharing of short-form travel videos on TikTok or Douyin. Current Issues in Tourism 25, 21 (2022), 3412–3424.

Yan Chen, Walter S Lasecki, and Tao Dong. 2021. Towards supporting programming education at scale via live streaming. Proceedings of the ACM on Human-Computer Interaction 4, CSCW3 (2021), 1–19.

Yaqi Zheng and Xipei Ren. 2022. Developing a Multimodal HMI Design Framework for Automotive Wellness in Autonomous Vehicles. Multimodal Technologies and Interaction 6, 9 (2022), 84.