# Automatic Generation of Road Trip Summary Video for Reminiscence and Entertainment using Dashcam Video

Kana Bito
Nagoya University
Nagoya, Aichi, Japan
bito.kana@g.sp.m.is.nagoya-u.ac.jp

Itiro Siio
Ochanomizu University
Tokyo, Japan
siio@acm.org

Yoshio Ishiguro
Nagoya University
Nagoya, Aichi, Japan
Tier IV, Inc.
Nagoya, Aichi, Japan
ishiy@acm.org

Kazuya Takeda
Nagoya University
Nagoya, Aichi, Japan
Tier IV, Inc.
Nagoya, Aichi, Japan
kazuya.takeda@nagoya-u.jp

## ABSTRACT

Vehicle dashboard cameras are becoming an increasingly popular kind of automotive accessory. While it is easy to obtain the high-definition video data recorded by dashcams using Secure Digital memory cards, this data is rarely used except for safety purposes because it takes substantial time and effort to review or edit many hours of such recorded videos. In this paper, we propose a new usage for this data through the automatic video editing system we have developed that can create enjoyable video summaries of road trips utilizing video and other data from the vehicle. We also report the results of comparisons between automatically edited videos created by the proposed system and manually edited videos created by study participants. The prototype developed in this study and the findings from our experiments will contribute to improving the driving experience by providing entertainment for automobile users after road trips, and by memorializing their travels.

## CCS CONCEPTS

• **Human-centered computing → Interface design prototyping**.

## KEYWORDS

dashcam, dashboard camera, video editing, video summarization, automatic video creation

**ACM Reference Format:**
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 (AutomotiveUI '21), September 9–14, 2021, Leeds, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3409118.3475151


*AutomotiveUI '21, September 9–14, 2021, Leeds, United Kingdom*

ACM ISBN 978-1-4503-8063-8/21/09…$15.00
https://doi.org/10.1145/3409118.3475151

## 1 INTRODUCTION

With the advent of smartphones and wearable cameras, it has become easy to record personal experiences as video[9, 14]. Widespread use of social networking sites has also provided more opportunities for sharing these videos with others. Automobiles have also been equipped with a large number of cameras and sensors, allowing the experience of riding in an automobile to be recorded in detail. In the future, as fully self-driving cars become a reality, sensing cameras are expected to become more sophisticated and powerful [15], but data from such sensors would likely be used mainly for autonomous driving navigation, driver monitoring, and traffic safety.

In the past few years, in-vehicle systems that display virtual games and information about nearby businesses such as restaurants, etc., which are updated with the movement of the vehicle, have been proposed as new applications for the video and other sensor information data obtained from vehicles [13, 24]. These applications are aimed at improving user experience (UX) during the trip, but very few studies have focused on after-travel UX. Since the timeframe of the user experience extends into the post-ride period [29], improving the after-travel UX can improve the UX of the overall transportation experience. Therefore, this study focuses on post-trip reflection and entertainment, intending to improve post-ride UX.

When developing the system described in this study, we focus on dashcams, which have rapidly become popular in recent years [16]. The primary application for dashcams is to record driving safety emergencies, such as traffic accidents and encounters with unsafe drivers, therefore the recorded data is usually of little use and is automatically discarded as it exceeds the capacity of the recording medium. Many recently manufactured dashcams are equipped with GPS and multiple HD cameras, allowing them to record high-quality video of the interior of the vehicle and exterior driving environment [21]. The recording video data is of sufficient quality that it can be used for home videos and for videos created for sharing on social networking sites.

There are many advantages to creating travel videos from data automatically recorded by a dashcam. First of all, there is no need to purchase or bring video equipment to record a video of your

trip, and steadier video can be recorded than when using a handheld video camera or smartphone. In addition, images recorded unconsciously (automatically) are more likely to be spontaneous and bring back memories of the past than images recorded consciously [23]. Therefore, it is thought that viewing a travelogue movie created from automatically recorded dashcam video is more likely to evoke memories of the trip. Furthermore, more than 50% of the conversation which occurs in the car is related in some way to the location around the vehicle [18], making it very easy to generate scenes with narration related to the location by simply adding the conversation in the car to the corresponding video.

On the other hand, the technology needed for handling large amounts of data is still under development. For example, editing dashcam video requires a lot of time and effort due to the volume of video recorded. As a result, although video data from a dashcam would be useful material for a travel movie, it would be difficult to manually review all of the recorded data, select the limited number of scenes that contain impressive events, merge these highlight scenes, add narration and add special effects. By using software such as Quick [10] or Magisto [28], we can automatically edit videos taken with action cameras by cutting, merging, adding effects, etc. However, even when using this kind of software, the user still needs to manually select the scenes to be included in the video, input captions or other text, and so on.

To address this problem, we have created a prototype of a system that automatically edits dashcam video from car trips, using dashcam data and three types of sensors. This paper's contributions are as follows:
1) We propose an automatic summarization system of road trip video, and explain the details of the prototype design;
2) We report the results of applying the suggestion system during five family trips and the comparison results between the videos automatically created by the proposed system and those manually created by study participants;
3) We discuss ways to utilize in-car videos and suggest possible improvements based on our observations.

## 2 RELATED WORK

### 2.1 Utilization of Dashcams

Many studies have investigated the use of dashcam data for traffic safety. For example, there have been studies on using dashcams for traffic accident prediction [3, 27], real-time detection of on-street parking detection [17], and automatic detection of distracted drivers [4]. A dashcam equipped with a system that detects high-risk behavior and incidents in real-time using built-in AI and G-force accelerometer data, and alerts the driver using speech commands, is now available [22]. The primary purpose of this study is to utilize dashcam data for post-travel entertainment, however, there may be additional applications, such as reviewing deliveries or taxi rides, for example.

### 2.2 Automatic Video Summarization

Many previous studies have focused on automatic video summarization. For example, video summarization techniques based on singular value decomposition (SVD) and clustering [7], or color feature extraction from video frames and k-means clustering algorithms [5], or which detect scene changes by modeling graphs [19] have all been proposed. In these studies, the summarization is mainly focused on scene changes, however when dealing with video of a car's interior, for example, there is little change in visual scenes because people do not move around much when traveling by car. Also, from a viewpoint of generating an attractive video summary, it is not necessary to cut out all scenes exhaustively, since the importance of driving scenes varies greatly depending on location and the surrounding driving environment. For example, scenes such as "approaching the destination" and "tourist attractions" are important, while scenes of expressways travel and residential areas may not important. Therefore, these previously proposed methods are not suitable for summarizing videos of a road trip.

### 2.3 Summarizing of Experiences and Memories

There have been many studies on summarizing experiences and memories using video of life events. For example, researchers have attempted to detect the startle response, which is a reaction that appears when a person is surprised, and to automatically record events that elicit this reaction during the user's daily life [11]. Another study proposed a method of automatically summarizing video recordings of daily life using an EEG and a wearable camera [1]. There is also an attempt to create video summaries of user experiences using video cameras, microphones, ID tags with infrared LEDs, and signal trackers [25]. From these studies, we can see that scenes that were considered to be highly important were those which involved user movement or an obvious physical reaction. However, the subjects in these studies were required to wear special sensors, and some of these proposed methods, if applied as proposed, could interfere with driving. Therefore, in this study, we designed a system that selects scenes at the moment of the user movement in a less intrusive manner, by using a seat-mounted pressure mat and a door-mounted magnetic reed switch. In addition, we extracted scenes that included lively conversation using the volume of recorded audio signals.

Many other methods have been proposed which use widely available devices for identifying and summarizing key life events. ComicDiary uses portable information terminals (PDAs) and kiosks to automatically generate comic-book-style diaries which include personal profiles, activity records, and recordings of interactions with other users while participating in academic conferences [26]. Video-Recording Your Life uses a GoPro wearable camera and the accelerometer of a smartphone to automatically extract scenes that users consider to be interesting from video recordings of their daily lives [2]. The wearable camera SenseCam camera used in Microsoft's "MyLifeBits" project incorporates several sensors to automatically takes pictures when it detects significant changes in light intensity in front of the camera, or changes in user body temperature, just by wearing it [6, 12]. The use of widely available components in such products makes it more likely that large numbers of people will use them. In this study, we use dashcams, which have rapidly increased in popularity in recent years, and can be easily purchased, to summarize the activity of users while traveling in an automobile.

The timeline function of Google Maps is one of the most widely used automatic travel recording systems in the world [8]. This application plots information such as travel routes, places visited and places where users stay on a map, and automatically summarize their travel records. It can also display travel photos stored in Google Photos. In this study, we also designed our system to automatically incorporate information such as travel route, places and times visited, and travel photos taken by the user into the video.

## 3 HARDWARE

The hardware used to create this prototype is shown in Figure 1. Each of these components is explained in detail below.

### 3.1 Dashcam

We use a Yupiteru Q-20P dashcam, which can shoot 360 degrees horizontally and 240 degrees vertically, and is equipped with a GPS positioning function. The acquired trip video and location history are stored on an SD card as MP4 and NMEA (National Marine Electronics Association) files, respectively. For both types of data, one file consists of approximately 1 minute of data (hereinafter referred to as a "1-minute file"), so a large number of files are generated during a trip. The MP4 files can be saved in two formats: fisheye or 2-split (the top half shows the area outside the vehicle and the bottom half shows the area inside the vehicle). We choose 2-split format for ease of data handling. GPS position information is recorded in the NMEA file every second. The dashcam is powered by the ACC (Accessory) power line of the car and stops recording when the ACC power is shut down. We judge that the vehicle had arrived at the resting point or destination when the interval of the NMEA timestamp exceeds 10 minutes.

### 3.2 In-vehicle Sensing

We installed sensors and switches to help record the situation inside the vehicle, as well as a PC (MacBook Pro Apple M1) and an Arduino to operate the system. The following sensors and switches were connected to the Arduino. To detect the opening and closing of the door, we installed a reed switch on the body of the car and a magnet on the door. We also installed a mat switch to detect when someone sat down in, or got up from, a seat, and two push-buttons to record video either inside or outside of the car, allowing users to manually capture scenes. In addition, we created software to record the sensor and switch information on a PC during the drive. When using the proposed automatic editing system, the use of a dashcam is mandatory, but the use of in-vehicle sensing data is optional.

## 4 AUTOMATIC EDITING SYSTEM

Our automatic video editing system creates a video summary of a road trip based on video files, log files of GPS, sensor and switch activity, obtained from the dashcam and Arduino. It also utilizes smartphone photos of the trip taken by the user. The automatic editing system is written in Python 3.7.9 and uses the FFmpeg command-line tool for video processing. The proposed system performs the following five processes; (1) adjusting the video size and brightness, (2) automatic detection and selection of highlight scenes, (3) adding user photos taken during the trip, (4) generating a progress map, and (5) merging the highlight videos and adding special effects. The details of each process are described below.

### 4.1 Adjusting Video Size and Brightness

*4.1.1 Adjusting video size.* The video data obtained from the dashcam used in this research is 2048 × 1536 pixels (width x height) in size, recorded at 28 frames per second. Each frame is divided into two parts, a top and a bottom image, which show scenes recorded outside and inside the vehicle, respectively. In order to use as much data from the recorded videos as possible, the 1360 × 765 pixels in the center of the upper and lower videos are copied when a scene is selected. The videos are then resized to 1280 × 720, which is the output size of the 720p HD video.

*4.1.2 Brightness Adjustment.* Since the primary purpose of the dashcam is to record videos of the area outside of the vehicle, the camera exposure is adjusted to the light level of the scenery outside. Therefore, videos of the inside of the car appear dark even in the daytime, and almost nothing can be seen at night. Therefore, the following processing was performed on videos of the inside of the car. First, gamma correction was performed, with gamma = 2.3 to brighten the entire video. Next, the video's saturation and contrast, which were also reduced using gamma correction were adjusted. The results of this processing are shown in Figure 2.

### 4.2 Automatic Detection and Selection of Highlight Scenes

Our system automatically extracts scenes of noteworthy events that occur during the trip (hereinafter referred to as "highlight scenes"). Based on a review of the related work, we believe that information about area information and the travel route are important for a travel summary. Therefore, we extract the time of departure and the time of arrival, as well as what we consider to be key points along the travel route, such as when the geographical area changes, or when passing near a landmark or the entrance or exit of a famous road that characterize the area. It can also be seen that capturing events occurring inside the car are also important. One type of scene in which there is a large amount of movement inside the car is the moment when the users open the door and sit down inside. Therefore, scenes of opening and closing the doors, as well as the moments when users are sitting down or getting up from their seats, were captured in this study. In addition, we assumed that the scenes, which included lively conversations were also important, and selected these scenes too. As described above, automatic editing is designed to extract the scenes necessary for summarizing a trip, but this may not include all the scenes the users feel are important. Therefore, we installed two push-button switches that allow the users to manually select any portion of the interior or exterior video as a highlight scene. The following is a detailed description of these highlighted scenes.

*4.2.1 Departure.* The time at which the vehicle's ACC power was turned on was assumed to be the time of departure. After the dashcam had started recording, the system extracted 7 seconds of exterior video and 10 seconds of interior video. The departure time obtained from the NMEA file was used to insert text and a synthetic

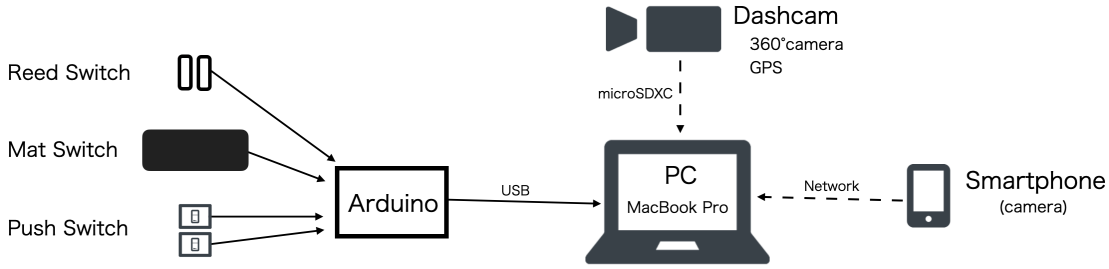


**Figure 1: Schematic of the hardware to be installed in the car.**

**Figure 2: Brightness adjustment of the image from inside the car. Before adjustment (left) and after adjustment (right).**

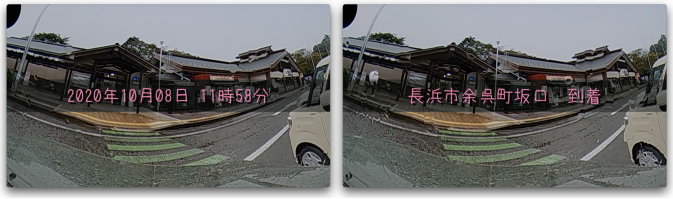


**Figure 3: Video of arrival at destination, automatically edited by the proposed system. First 3 seconds after arrival (left) and second 3 seconds after arrival (right).**

voice message into the exterior video scene. The gTTS (Google Text-to-Speech) library in Python was used to insert the synthetic voice.

*4.2.2 Arrival.* The time and address of the arrival location were obtained using NMEA files and a location information API (Yahoo! Japan's Yahoo! Open Local Platform). Text messages and synthetic voice descriptions were inserted into the exterior video, as shown in Figure 3.

*4.2.3 Changes in geographical area.* In this study, we extracted scenes that included the crossing of a prefectural border. We obtained the names of the prefectures using the coordinate information obtained from the NMEA file and the location information API. The location of the prefectural border was determined based on the changes in the acquired prefecture name, and exterior video was extracted from the moment of crossing the border for a period of 3 seconds. The name of the prefecture was shown with inserted text and was announced by a synthetic voice, as shown in Figure 4.

*4.2.4 Landmarks.* Using coordinate information obtained from the NMEA file and location information from the Yahoo! API, we obtained regional landmark information such as the names of large facilities (e.g., amusement parks, sports stadiums), sightseeing spots, and local place names. Based on the acquired information, we determined when the vehicle entered the vicinity of a landmark and extracted the following 3 seconds of exterior video. Place name information was displayed by inserting text and a synthetic voice message into the video, as shown in Figure 4. In this study, the vicinity of the landmark was extracted only when the confidence score assigned by the location information API to the acquired regional information was 99.9% or higher.

*4.2.5 Expressway entrances and exits.* Expressways were identified using vehicle speed information obtained from the NMEA file, and the average speed for one minute was calculated. In Japan, where this prototype was tested, general roads are defined as those with speed limits of less than 60km/h, while expressways are usually traveled at 60km/h or more. In addition, the Electronic Toll Collection (ETC) gates at the entrance and exit of an expressway require drivers to slow down to 20km/h or less when transiting. Therefore, in areas where there were intersections of expressways (the vehicle's speed was above 60km/h) and public roads (the vehicle's speed was below 60km/h), locations where the vehicle slowed to between 10km/h and 20km/h were considered to be the entrance or exit of an expressway. In these areas, we clipped the exterior video for 4 seconds, from 5 seconds before reaching minimum speed to 1 second before. In this study, we used speed changes to detect scenes of expressway travel, but we believe it would also be possible to obtain navigation map information and use it to identify and extract entrances to expressways and famous streets.

*4.2.6 Lively conversations.* Scenes that featured lively conversations were identified using audio data, and video of the interior of the car was then selected. A band-pass filter was applied to the audio data obtained from the MP4 file to extract the 300-3400 Hz frequency band of normal Japanese speech, and the video was extracted for 6 seconds, from 3 seconds before to 3 seconds after the amplitude of its audio data was above a certain threshold. We also used Python's SpeechRecognition library to display the subtitles of the conversations, as shown in Figure 4.

*4.2.7 Opening/closing of the door.* A reed switch installed on the door detected the opening and closing of the door, and the Arduino recorded the detection time. The video recorded by the dashcam was extracted for 4 seconds, from 2 seconds before to 2 seconds after detecting the opening and closing of the door.


**Figure 4: Highlight scenes automatically extracted and edited by the proposed system. Crossing a prefectural border (left), passing near a landmark (center), and lively conversation (right).**

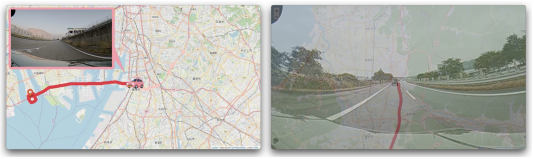

**Figure 5: Scenes that were not the target of highlight. Visualization of the route already traveled on a map, with current exterior scene (left), alpha-blending a map with a 100 × actual speed video (right).**

*4.2.8 When passengers occupy or vacate their seats.* A mat switch installed on the seat detected when a seat was occupied or vacated, and the Arduino recorded the detection time. The interior video was then extracted for 4 seconds before and after the detected time.

*4.2.9 When a manual record button is pressed.* The Arduino records the time when an interior scene or exterior scene manual record button is pressed. While the button is being pressed, the corresponding section of the video is extracted.

## 4.3 Adding Photos Taken During the Trip

Exif (Exchangeable image file format) information, including the date, time, and location of the photo, is recorded when photos are taken with smartphones or digital cameras. Therefore, if Exif information is embedded in a photo taken on a trip, the date and time when the photo was taken are retrieved from the photo's Exif information and the photo is displayed at the appropriate position in the travel video. The display time of such images was set to 1.2 seconds, and in the case of vertical photos, margins were added on the left and right to match the 16:9 ratio.

## 4.4 Generating a Map

When traveling by car, geographic information about the route and the current location of the vehicle are also important. The recorded 1-minute files that were not extracted highlight scenes are displayed on a map using Python's Folium library to visualize the acquired coordinates. A red line was used to indicate the path taken from the time of departure to the time of arrival at the destination, and a red goal marker was plotted when the destination was reached. When a trip is resumed, roads traveled so far are changed from red lines to pink lines, and roads traveled from the resumption of travel to the time of arrival at the next destination are replotted with red lines, as shown in Figure 5. In order to combine this map with exterior images, two types of representation methods were adopted. When using these representation methods, the first frame of each 1-minute file is cut out, and the absolute sum of the differences in RGB values of this frame and the corresponding pixels of the first frame of the previous 1-minute file is calculated. The following two processes are then performed by comparing this sum with a set threshold value.

If the difference is greater than the threshold, it is assumed that the change in the scenery outside the vehicle is large. A map with the travel route, a still image of the exterior video, and an illustration of the car is created for each minute of recorded data, and this map is displayed for about 0.3 seconds. Frame per second for the image of the exterior video is dropped to one frame per second. An example is shown in Figure 5. Hereinafter referred to as "map-based video".

Conversely, if the absolute sum of the difference in RGB values is less than the threshold, the change in scenery is assumed to be small. The exterior video cut out from the 1-minute file is sped up 100 times faster than normal, and alpha-blended with a map showing travel routes, as shown in Figure 5. Hereinafter referred to as "accelerated video".

## 4.5 Merging Videos and Adding Effects

A travelogue movie can then be created by connecting these files in chronological order. However, additional steps must be taken to avoid duplicating the same highlight scenes. In addition, if we connect the created highlight scenes and travel maps as they are, the video will be difficult to watch because the image on the screen will change too frequently. Therefore, the following process was applied to solve these problems.

*4.5.1 Processing order.* We prepared a two-dimensional array corresponding in size to the number of 1-minute files, and record the processing to be used for each 1-minute file using the numerical values corresponding to that processing. The order of the processes and the numbers corresponding to each highlight scene are as shown in Figure 6. For example, to create a departure highlight scenes from the 10th 1-minute file, 3 is stored in the 10th array. Multiple highlight scenes can be generated from the same 1-minute file, but to avoid duplication in the generated video, if the cut range and time are the same, only the highlighted scenes in the earliest processing order will be saved.

*4.5.2 Adjusting the map display method.* As explained earlier, when there is no highlight scene to display, the exterior image synthesized into the map is switched between a 100 x actual speed video and a 1-second frame drop according to the magnitude of the changes occurring in the outside scenery. However, excessive switching results in a video that is difficult to watch. Therefore, the following low-pass filter processing is applied to the array that specifies the processing for the 1-minute video, in order to suppress the change. The processing priority is: (1) > (2) > (3).

(1) When the map-based video is sandwiched between the 100 x normal speed videos, the accelerated video is displayed instead of the map-based video. For example:[ ..., 0, 1, 0, ... ] → [ ..., 0, 0, 0, ... ]

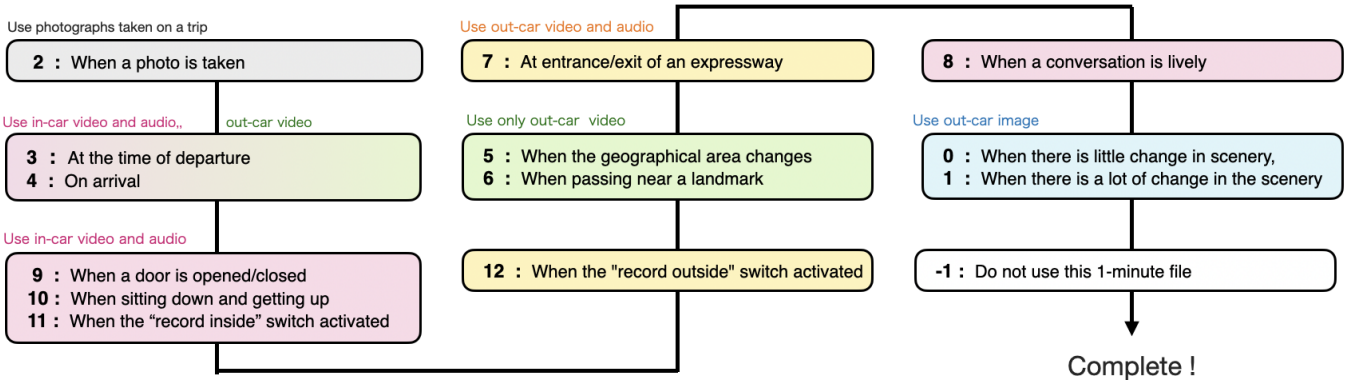


**Figure 6: Order of processing and priority of highlight scenes, numbers corresponding to each process.**

(2) If there are less than three consecutive accelerated videos, change all of the accelerated video into map-based videos. For example: [ ..., 1, 0, 0, 1, ... ] → [ ..., 1, 1, 1, 1, ... ]
(3) If the map-based video does not continue to be displayed, the map information is not used. For example:[ ..., 0, 1, 5, ... ] → [ ..., 0, -1, 5, ... ]

*4.5.3 Transitional effects.* The following visual and sound effects were applied when applicable.

**Speech balloons:** To smoothly switch between the map information (when the number stored in the array is 1) and the video (when the number is not 1), a speech balloon effect was used. Specifically, when switching from the map information to a moving image, the balloon displaying the image of the outside of the car plotted on the map section is gradually enlarged to make the switch. Conversely, when switching from video to map information, the full-size image of the car exterior is gradually reduced to the size of the balloon plotted on the map information, as shown in Figure 7.

**Fade-in/fade-out:** There is a time gap between the arrival at a destination and the next departure because data is not recorded while the vehicle's ACC power is off. Therefore, the screen was darkened by adding a fade-in process at the time of departure and a fade-out process at the time of arrival to indicate the arrival at a destination.

**Radial wipe effect:** When switching from a normal speed video to a 100 x normal speed video, the playback speed changes significantly. Therefore, we added a radial wipe effect to smoothly connect these videos, as shown in Figure 8.

**Jingle sounds:** Sound effects were added to balloon effect transitions, radial wipe effects, and highlight scenes crossing the border of a prefecture and traveling in the vicinity of landmarks. Accelerated video is accompanied by an up-tempo jingle and map-based video is accompanied by a medium-tempo jingle.

*4.5.4 Merging video files.* All of the videos created using the process described above are given a name corresponding to the time during the journey when the scene was selected, and these videos are saved in the folder used for the completed video. For example, if a clip from the 30-second point in the 80th file is selected, the file is named 00000080_30.mp4. After all of the processing is completed, the processed videos in the folder for finished videos are sorted by name, then all of the files are concatenated to form a single video. Currently, the length of the video generated by the system corresponds to the length of the trip.

# 5 EXPERIMENT

## 5.1 Experimental Procedure

We applied our system to five family trips taken by one of the authors with her mother. During three of these trips, an experiment was conducted to compare the contents of the videos generated by the proposed system with those of manually edited videos. The PC used for the experiment was the same PC that was used for the in-vehicle system. The subjects who manually edited the videos were the mother who participated in the trip (P1) and a family member who did not participate in the trip (P2).

P1: Female in her 50s, with no experience editing videos
P2: Male in his 20s, with no experience editing videos

*5.1.1 Automatic video editing system.* The procedure for using the automatic video editing system is as follows. Since the recording is done by the dashcam and the system automatically edits the data, there is little work for the user to do.

(1) Start the dashcam and begin driving. Press the switch for adding highlight scenes manually when desired.
(2) After the trip, import the SD card data from the dashcam and the photos taken during the trip using smartphones or digital cameras into the PC that is running the system.
(3) Start the automatic video editing program.

In the current version of our prototype, it is also necessary for the user to start the software for acquiring the log of sensors and switches stored in the on-board PC at the beginning of the trip, but we plan to automate this process in future versions of our application.

*5.1.2 Manual editing by study participants.* For the videos manually edited by the study participants, the experiments were conducted in the following order, to avoid influencing the participants with the automatically edited video:

(1) Collect raw dashcam video during three road trips.
(2) Have the participant manually edit the videos by selecting highlight scenes and adding effects using iMovie software.

**Figure 7: Transitions with balloon effects (Zoom-in).**

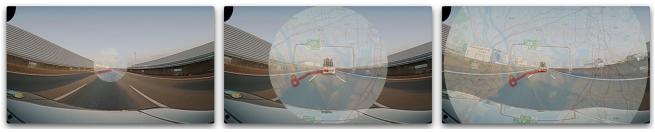

**Figure 8: Radiant wipe effect.**

(3) Show the participant the videos automatically generated by the proposed system.
(4) Interview the participant to obtain their opinions on the manually edited and automatically edited videos.

## 5.2 Results

Details of the five road trips are shown in Table 1. For three of these trips (Trips 1, 2, and 3) only data from the dashcam was used, and for the remaining two trips (Trip 4 and 5) video recorded of the interior and exterior of the vehicle using optional sensor and switches was also used. The dashcam was installed on the right-side rearview mirror and recorded the exterior and interior (driver and front passenger seats) of the car. The data was edited after returning home. The details and results of manual editing by subjects P1 and P2 for Trips 1, 4 and 5 are shown in Tables 2 and 3, respectively.

## 5.3 Automatic vs. Manual Editing

Using the method and data (Trip 1, 4, and 5) described in the previous section, we compared the results of automatic video editing by the proposed system with manual editing of the video by participants P1 and P2.

*5.3.1 Method.* F-measure is used in our evaluation of matching rate because it can comprehensively evaluate accuracy as it is the harmonic mean of precision and recall. There are two possible measurements of the match rate: overlapping time and overlapping scene. In this study, we adopted the consistency rate of scenes, and the average time per scene was compared separately. The definition of a single scene is as follows. During manual editing, we defined a scene as one that the editor selected with the intention of making it a single scene. Therefore, even if the combined videos were not consecutive in time, it was considered to be one scene if the content or subject of the video segments was the same. However, even within a single scene, if one editor selected a scene that was omitted during manual editing by another editor, it was considered not to be a match. During automatic editing by the system, all of the video selected during one highlight scene selection was considered to be one scene.

*5.3.2 Cost performance.* The time required for editing the travel summary video was about 181-283% of the driving time for P1 and about 144-209% of the driving time for P2, while the automated system took only about 23-25% of the driving time to generate a video, even though the number of highlight scenes and special effects (captions or other text and transitions) used by the automated system was almost the same or more than those used during manual editing.

The length of the generated videos was P1 > P2 > system for all of the three trips compared (Trips 1, 4, and 5). This can be attributed to the fact that the average length per highlight scene selected by the automated system was shorter than those selected by the study participants. The automated system focused on the tempo of the video, so the average length per scene was about 6-9 seconds, whereas the human editors were more aware of the flow of the conversation, so the average length per scene was much longer, about 40-45 seconds for P1 and about 15-25 seconds for P2.

*5.3.3 Consistency rate.* As a result, the consistency rate between automatically edited scenes and P1 edited scenes was about 19 to 49%, between automatically edited and P2 edited scenes about 24 to 31%, and between scenes manually edited by P1 and P2 about 44 to 45%. Thus, the consistency rate was low, i.e., less than 50%, for all of the comparisons.

## 5.4 Interview

We conducted follow-up interviews with the two participants who created manually edited travel videos. Table 4 were common opinions between P1 and P2.

# 6 DISCUSSION

## 6.1 Strengths and Weaknesses of Automatic and Manual Editing

One possible reason for the low correlation between the manually and automatically edited videos is that they have a different focus and different strengths and weaknesses, i.e., different scenes were considered to be suitable as highlights by our system than were selected by our human editors. For example, one type of scene that was seldom selected during manual editing was exterior video recorded near landmarks. Therefore, if we program the system not to select scenes near landmarks, the rate of agreement between the automatically edited and manually edited video would increase. However, some of the scenes selected by the system were difficult to select manually due to difficulty handling the data, so increasing the correlation rate between the selected scenes would not necessarily result in the generation of better travel summary videos. In addition, both P1 and P2 expressed the opinion that selecting scenes near landmarks would be desirable. Based on the results of the above experiments and interviews, we have summarized the merits of both automatic and manual editing and their suitable uses as follows.

The advantages of the proposed system are that it requires less time and effort, and that it can present map information and surrounding area information that manual editing cannot. Therefore, the automated system is particularly useful when traveling for a long time, and is suitable for creating travel diaries focusing on travel routes. P2 said, *"The system can create a nice-looking diary with little effort. The tempo is also good, so it is suitable for sharing on social media"*.

**Table 1: Trip and video editing detail when using the automated video editing system.**

| Trip # | Length of trip | Driving time | Use of switches | Processing time | Length of generated video | # of highlight scenes | # of effects |
|---|---|---|---|---|---|---|---|
| **1** | **8h 54m** | **2h 39m** | **No** | **37m 21s** | **4m 43s** | **24** | **58** |
| 2 | 1d 12h 23m | 10h 38m | No | 2h 54m 31s | 15m 24s | 84 | 206 |
| 3 | 1d 11h 40m | 14h 45m | No | 4h 22m 32s | 19m 38s | 116 | 328 |
| **4** | **5h 18m** | **1h 38m** | **Yes** | **24m 26s** | **5m 05s** | **43** | **54** |
| **5** | **8h 24m** | **1h 48m** | **Yes** | **26m 08s** | **3m 33s** | **21** | **39** |

**Table 2: Details of manual editing by P1 and resulting video.**

| Trip # | Editing time (Scene selection + Adding effects) | Length of generated video | # of highlight scene | # of effects |
|---|---|---|---|---|
| 1 | 3h 50m 11s ( 2h 23m 32s + 1h 27m 39s ) | 8m 10s | 11 | 16 |
| 4 | 4h 47m 35s ( 3h 42m 24s + 1h 05m 11s ) | 15m 03s | 21 | 41 |
| 5 | 4h 37m 03s ( 2h 28m 25s + 2h 08m 38s ) | 16m 27s | 23 | 45 |

**Table 3: Details of manual editing by P2 and resulting video.**

| Trip # | Editing time (Scene selection + Adding effects) | Length of generated video | # of highlight scenes | # of effects |
|---|---|---|---|---|
| 1 | 3h 45m 17s ( 2h 50m 08s + 55m 09s ) | 5m 52s | 20 | 38 |
| 4 | 3h 48m 25s ( 3h 27m 34s + 20m 51s ) | 9m 02s | 34 | 28 |
| 5 | 3h 08m 10s ( 2h 38m 59s + 29m 11s ) | 8m 31s | 20 | 22 |

**Table 4: Common opinions between P1 and P2.**

| | |
|---|---|
| Q. How did you feel about manually editing the videos? | • Manual editing, especially adding text, is difficult and hassle.<br>• I can edit manually for about an hour, but longer than that is a pain.<br>• The scenery outside was more boring than I expected, so my video was all about what happened inside of the car.<br>• It was more of daily conversation than travel conversation. |
| Q. What do you think of the videos automatically generated by the system? | • I can tell that they are having fun and I would love to use it when it is released as an app.<br>• I like that there are subtitles.<br>• It's good that there is a map, so you can see where you've driven and how you traveled.<br>• The interesting parts of the conversation were not cut out. |
| Q. What would you like to be able to see or do when editing manually? | • What I would like to do manually is to select the interesting parts. |

The advantages of manual editing are that it allows the editing of conversations using common sense and a proper understanding of the situation. Therefore, manual editing would be more suitable for creating memory videos focusing on the content of conversations and the relationships between the travelers, where its main use would be as a home video to be shared among family members and relatives.

## 6.2 Lessons for Automated Editing of Dashcam Travel Videos

**Fully automated editing (video focusing on travel routes):** Follow-up studies should consider the following points. In this study, the selection of scenes with lively conversations is determined by the volume threshold, so this prototype system may include undesirable scenes such as arguments or loud environmental sounds. For example, the introduction of facial recognition and laughter recognition would solve this problem. In addition, head

tracking or eye tracking would enable us to recognize which landmarks are important and to cut out scenes without clear landmarks (e.g., ocean or mountain views). Furthermore, since recollection and remembering are enhanced by negative affect, positive affect, and arousal stimuli [20], we believe that we can generate summary videos that are more suitable for reflecting on memories by recognizing these emotions and automatically cutting them out. In the prototype, we focus on the tempo of the video and have not sufficiently considered the cutting time for each highlight scene. We will also need to find the appropriate duration for each highlight scene.

**Fully automated editing (video focusing on conversations):** The correlation rate between the videos manually edited by the two participants was below 50% in all three experiments. The likely reason for this was that the scenes that each person wanted to include in the video were different. When asked what was important when editing the video, P1 said *"to include funny remarks and interesting conversations"*, while P2 said, *"to understand the flow of the trip"*. Based on these remarks, a future configuration of the system could include the following steps:

(1) Dividing the video into scenes based on the content of the conversation and activities in the car.

(2) Categorization of the scenes (e.g., scenes of eating and drinking, talking about travel, talking about family, etc.)

(3) Picking out scenes based on user preferences.

(4) Connecting the selected scenes with visual and sound effects.

However, when classifying scenes, there are issues such as the possibility that the parts of scenes that users find interesting may differ from person to person. For example, conversations that only the passengers understand, conversations that only family members understand, etc.

**Semi-automatic editing:** As a result of our interviews with the participants, we found that long hours of manual video editing are a pain. However, one noted that it was interesting to look back on the trip if it is for a short time (P1). Manual video editing will be more enjoyable if you can complete it in a few tens of minutes, even if you are traveling for a long time. It will be essential to design a system that allows users to edit the video without having to review the entire corpus of video data, by either dividing the video into scenes and only allowing users to choose whether or not to keep the suggested scenes, or by indexing the video. We also believe it is important to automate the less cost-effective processes such as size and brightness adjustments.

## 6.3 Considerations for video editing systems in general

Before uploading these travelogue movies to social media sites, it would be desirable to apply mosaic processing to other people's faces and car license plates for privacy reasons. It might be also helpful to allow users to designate the desired length of the video they would like to generate. In addition, since the video recorded at night can be quite hard to see even after correcting the brightness, some additional processing would be required.

## 6.4 Limitations

For the experiment, we used short-distance (within the prefecture), medium-distance (in a neighboring prefecture), and long-distance (in a remote prefecture) travel data. These data included a variety of weather conditions and times. They also included travel data from different regions, such as expressways, natural countryside, and residential areas. Therefore, we were able to try out a variety of patterns. Also, by having both the passenger and a third party do the manual editing, we were able to get different perspectives. Therefore, we believe that the results of this study are a useful first step for future experiments and concepts. However, It's not an exhaustive survey, since we only tested it on a parent-child trip with two people and the sample size was small. In the future, it is necessary to conduct a larger-scale follow-up survey.

In addition, scenes selected using the manual record switches were not always included in the manually edited travel videos. The reason given for this was, *"I thought it was interesting at the time, but in retrospect, I did not find it very interesting"*(P1). During our experiment, there were few opportunities to activate in-vehicle sensors, so we were not able to fully investigate the use of these switches. Further follow-up study is required.

# 7 CONCLUSION & FUTURE WORK

In this study, we have described a system we developed that can automatically select video highlights of an automobile trip from data recorded with a dashcam, and generate a travel video summary. After comparing videos generated automatically by the proposed system with those created through manual editing, we found advantages and disadvantages to each approach, and conclude that it is necessary to consider the purpose of the video being created when deciding the configuration of the summarization system.

As automated driving technology becomes more mature, we believe that we will be able to provide a much better post-travel user experience than is possible using currently available methods such as dashcams, by creating a video summarization system that uses the many high-quality cameras and sensors installed in automated vehicles.

# REFERENCES

[1] Kiyoharu Aizawa, Ken-Ichiro Ishijima, and Makoto Shiina. 2001. Summarizing wearable video. In *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, Vol. 3. IEEE, 398–401. https://doi.org/10.1109/ICIP.2001.958135

[2] Daniel Buschek, Michael Spitzer, and Florian Alt. 2015. Video-Recording Your Life: User Perception and Experiences. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA '15)*. Association for Computing Machinery, New York, NY, USA, 2223–2228. https://doi.org/10.1145/2702613.2732743

[3] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. 2017. Anticipating Accidents in Dashcam Videos. In *Computer Vision – ACCV 2016*, Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (Eds.). Springer International Publishing, Cham, 136–153. https://doi.org/10.1007/978-3-319-54190-7_9

[4] Jaco Cronje and Andries P. Engelbrecht. 2017. Training Convolutional Neural Networks with Class Based Data Augmentation for Detecting Distracted Drivers. In *Proceedings of the 9th International Conference on Computer and Automation Engineering* (Sydney, Australia) *(ICCAE '17)*. Association for Computing Machinery, New York, NY, USA, 126–130. https://doi.org/10.1145/3057039.3057070

[5] Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes, Antonio da Luz, and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters* 32, 1 (2011), 56–68. https://doi.org/10.1016/j.patrec.2010.08.004 Image Processing, Computer Vision and Pattern Recognition in Latin America.

[6] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong. 2002. MyLifeBits: Fulfilling the Memex Vision. In *Proceedings of the Tenth ACM International Conference on Multimedia* (Juan-les-Pins, France) *(MULTIMEDIA '02)*. Association for Computing Machinery, New York, NY, USA, 235–238. https://doi.org/10.1145/641007.641053

[7] Yihong Gong and Xin Liu. 2000. Video summarization using singular value decomposition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, Vol. 2. IEEE, 174–180. https://doi.org/10.1109/CVPR.2000.854772

[8] Google. [n.d.]. Google Maps Timeline. Retrieved July 18, 2021 from https://www.google.com/maps/timeline

[9] GoPro. 2020. HERO9 Black. Retrieved July 18, 2021 from https://gopro.com/en/us/shop/cameras

[10] GoPro. 2021. GoPro Quik version 8.9.1. Retrieved July 18, 2021 from https://gopro.com/en/us/shop/quik-app-video-photo-editor

[11] Jennifer Healey and Rosalind W Picard. 1998. StartleCam: a cybernetic wearable camera. In *Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215)*. IEEE, 42–49. https://doi.org/10.1109/ISWC.1998.729528

[12] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. 2006. SenseCam: A Retrospective Memory Aid. In *Proceedings of the 8th International Conference on Ubiquitous Computing* (Orange County, CA) *(UbiComp'06)*. Springer-Verlag, Berlin, Heidelberg, 177–193. https://doi.org/10.1007/11853565_11

[13] Holoride. 2019. Holoride. Retrieved July 18, 2021 from https://www.holoride.com

[14] Insta360. 2021. Insta360 Go 2. Retrieved July 18, 2021 from https://www.insta360.com

[15] Yano Research Institute. 2020. <Smart City> Key Device Components for ADAS/AD 2020. Retrieved July 18, 2021 from https://www.yanoresearch.com/market_reports/C61119900

[16] Mordor Intelligence. [n.d.]. DASHBOARD CAMERA MARKET - GROWTH, TRENDS, COVID-19 IMPACT, AND FORECASTS (2021 - 2026). Retrieved July 18, 2021 from https://www.mordorintelligence.com/industry-reports/dashboard-camera-market

[17] Akihiro Matsuda, Tomokazu Matsui, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. 2021. A Method for Detecting Street Parking Using Dashboard Camera Videos. *Sensors and Materials* 33, 1 (2021), 17–34. https://doi.org/10.18494/SAM.2021.2998

[18] Kohei Matsumura and Yasuyuki Sumi. 2014. What Are You Talking About While Driving? An Analysis of In-Car Conversations Aimed at Conversation Sharing. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Seattle, WA, USA) *(AutomotiveUI '14)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2667317.2667417

[19] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. 2003. Automatic video summarization by graph modeling. In *Proceedings Ninth IEEE International Conference on Computer Vision*, Vol. 1. IEEE, 104–109. https://doi.org/10.1109/ICCV.2003.1238320

[20] Kevin N. Ochsner. 2000. Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology: General* 129, 2 (2000), 242–261. https://doi.org/10.1037//0096-3445.129.2.242

[21] Grand View Research. 2020. Dashboard Camera Market Size, Share & Trends Analysis Report By Technology (Basic, Advanced, Smart), By Product, By Video Quality, By Application, By Distribution Channel, By Region, And Segment Forecasts, 2020 - 2027.

[22] Samsara. [n.d.]. AI Dash Cams. Retrieved July 18, 2021 from https://www.samsara.com/products/safety/dash-cam

[23] Abigail J. Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. 2007. Do Life-Logging Technologies Support Memory for the Past? An Experimental Study Using Sensecam. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/1240624.1240636

[24] Honda Developer Studio. 2019. Honda Dream Drive. Retrieved July 18, 2021 from https://developer.hondainnovations.com

[25] Yasuyuki Sumi, Sadanori Ito, Tetsuya Matsuguchi, Sidney Fels, Shoichiro Iwasawa, Kenji Mase, Kiyoshi Kogure, and Norihiro Hagita. 2007. Collaborative Capturing, Interpreting, and Sharing of Experiences. *Personal Ubiquitous Comput.* 11, 4 (April 2007), 265–271. https://doi.org/10.1007/s00779-006-0088-1

[26] Yasuyuki Sumi, Ryuuki Sakamoto, Keiko Nakao, and Kenji Mase. 2002. ComicDiary: Representing Individual Experiences in a Comics Style. In *Proceedings of the 4th International Conference on Ubiquitous Computing* (Göteborg, Sweden) *(UbiComp '02)*. Springer-Verlag, Berlin, Heidelberg, 16–32. https://doi.org/10.1007/3-540-45809-3_2

[27] Yoshiaki Takimoto, Yusuke Tanaka, Takeshi Kurashima, Shuhei Yamamoto, Maya Okawa, and Hiroyuki Toda. 2019. Predicting Traffic Accidents with Event Recorder Data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility* (Chicago, IL, USA) *(PredictGIS'19)*. Association for Computing Machinery, New York, NY, USA, 11–14. https://doi.org/10.1145/3356995.3364535

[28] Vimeo. 2021. Magisto Video Editor & Maker version 6.6.1. Retrieved July 18, 2021 from https://www.magisto.com

[29] Roto Virpi, Law Effie, Vermeeren Arnold, and Hoonhout Jettie. 2011. UX-WhitePaper. Retrieved July 18, 2021 from http://www.allaboutux.org/files/UX-WhitePaper.pdf