Chapter 14
Content-Based Driving Scene Retrieval Using Driving Behavior and Environmental Driving Signals

Yiyang Li, Ryo Nakagawa, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda

Abstract With the increasing presence of drive recorders and advances in their technology, a large variety of driving data, including video images and sensor signals such as vehicle velocity and acceleration, can be continuously recorded and stored. Although these advances may contribute to traffic safety, the increasing amount of driving data complicates retrieval of desired information from large databases. One of our previous research projects focused on a browsing and retrieval system for driving scenes using driving behavior signals. In order to further its development, in this chapter we propose two driving scene retrieval systems. The first system also measures similarities between driving behavior signals. Experimental results show that a retrieval accuracy of more than 95 % is achieved for driving scenes involving stops, starts, and right and left turns. However, the accuracy is relatively lower for driving scenes of right and left lane changes and going up and down hills. The second system measures similarities between environmental driving signals, focusing on surrounding vehicles and driving road configuration. A subjective score from 1 to 5 is used to indicate retrieval performance, where a score of 1 means that the retrieved scene is completely dissimilar from the query scene and a score of 5 means that they are exactly the same. In a driving scene retrieval experiment, an average score of more than 3.21 is achieved for queries of driving scenes categorized as straight, curve, lane change, and traffic jam, when data from both road configuration and surroundings are employed.

Keywords Content-based retrieval • Driving data • Drive recorder • Similarity measure • Surrounding environment

Y. Li (✉) • R. Nakagawa • C. Miyajima • N. Kitaoka • K. Takeda
Graduate School of Information Science, Nagoya University, Nagoya, Japan
e-mail: yiyang.li@g.sp.m.is.nagoya-u.ac.jp

G. Schmidt et al. (eds.), *Smart Mobile In-Vehicle Systems: Next Generation Advancements*, DOI 10.1007/978-1-4614-9120-0_14,

14.1 Introduction

Drive recorders are used to investigate the causes of traffic accidents and to improve drivers' safety awareness. With the increasing presence of more advanced drive recorders, a large variety of driving data, including video images and sensor signals such as vehicle velocity and acceleration, can be continuously recorded and stored. Although these advances may contribute to traffic safety, the increasing amount of driving data complicates retrieval of desired information from large databases. Some researchers have studied methods for recognizing driving events, such as lane changing and passing, using HMM-based dynamic models [1–3]. In our previous work, a similarity-based retrieval system for finding driving data was proposed [4]. However, since our method used differences in histograms of driving behavior signals as the similarity measurement, it did not efficiently use dynamic information from driving scenes for retrieval. In this chapter, we study two driving scene retrieval systems that utilize dynamic information from driving scenes.

In the first study, we focus on driving behavior signals. The first retrieval system captures dynamic information from driving scenes by directly using sequences of driving behavior signals and utilizes changes in these signals over time. Six kinds of driving behavior signals (velocity, longitudinal and lateral acceleration, gas and brake pedal pressures, and steering angle) are used for calculating similarity between driving scenes. We compared the use of both early and late integration to integrate these signals.

In the second study, we focus on environmental driving data that is collected from the road and surrounding vehicles. The second retrieval system uses a similarity measure to compare the road configuration and motion of surrounding vehicles. Positions of surrounding vehicles and roadside barriers are detected with laser scanners mounted on the front and back of an instrumented vehicle, and the velocities of surrounding vehicles are estimated from their relative positions to the vehicle. Each scanned frame of a driving scene is categorized based on three general features, i.e., road type, congestion level, and the positions of surrounding objects. Also, the motion of each surrounding vehicle is tracked to obtain its motion features, so we can measure the similarity between vehicles. Categorization results and detected vehicle path are integrated to measure similarity between driving scenes.

14.2 Data Collection

The driving data used in our study was collected on real roads and was recorded using the instrumented vehicle shown in Fig. 14.1. The collected signals included velocity [km/h], longitudinal and lateral acceleration [G], gas and brake pedal pressures [N], and steering wheel angle [deg]. Two laser scanners were mounted on the front and back of the vehicle to detect surrounding objects. The laser scanners covered 80° arcs at both the front and back of the vehicle, to an effective

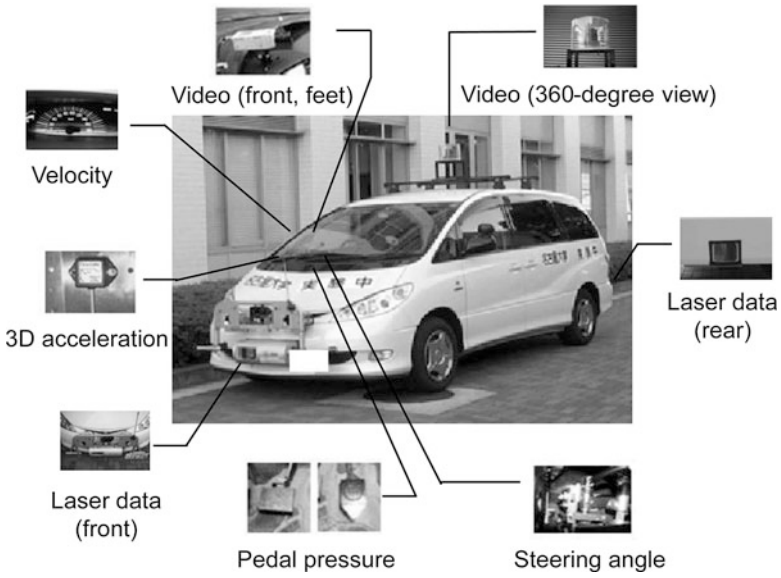


Fig. 14.1 Instrumented vehicle used for driving data collection [8]

range of about 100 m to the front and 55 m to the rear. A Kalman filter was employed to predict the motions of objects in blind areas. In order to assist in the subjective confirmation of retrieved scenes, synchronously recorded front and driver's feet scenes, as well as a 360° panoramic scene of the surroundings from an omnidirectional camera, were also available for every retrieved scene.

14.3 Driving Scene Retrieval Using Driving Behavior Signals

In this section, we describe the first similarity-based driving scene retrieval system, which uses similarity of driving behavior signals. Six driving signals (velocity, longitudinal and lateral acceleration, gas and brake pedal pressures, and steering angle) were used for calculating similarity between driving scenes. We compared the use of early and late integration to integrate these signals.

14.3.1 Integration Methods for Driving Behavior Signals

14.3.1.1 Method 1: Early Integration

We retrieved similar driving scenes using two methods, early and late integration. Figure 14.2 shows the procedure for early integration. The six kinds of signals mentioned above were extracted from the scene to be retrieved, and each signal

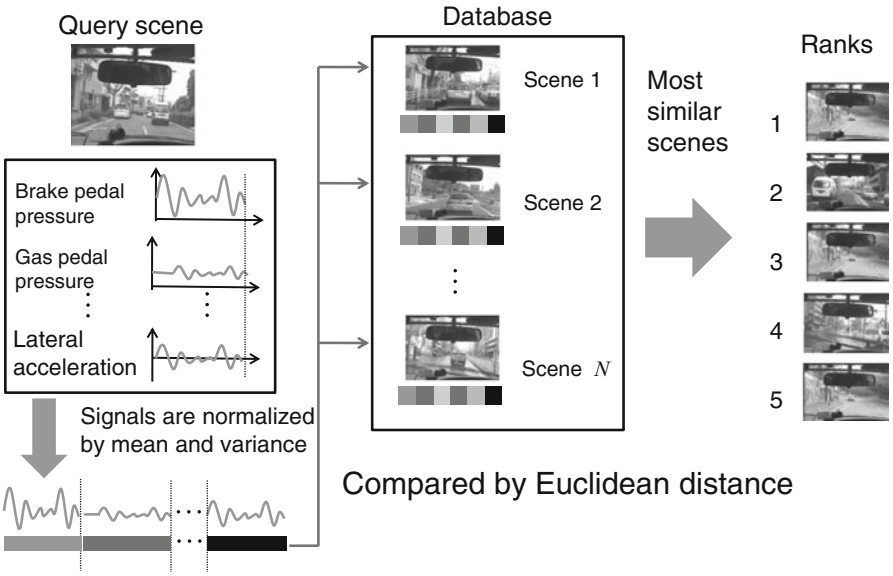


Fig. 14.2 Driving scene retrieval using driving behavior signals (early integration)

was normalized by mean and variance using all the data for all the drivers. The normalized signals of the query scene were represented as a vector, and the Euclidean distance between the vectors of the query scene and every scene in the database was measured. The database for the search consisted of about 200,000 vectors, one for each recorded scene. A fast retrieval technique was used to reduce retrieval time. The top five scenes with the smallest distances were chosen as similar scenes.

14.3.1.2 Method 2: Late Integration

The other retrieval method used was late integration, shown in Fig. 14.3. Each of the six kinds of signals of a scene was represented as a vector, and the Euclidean distance between the vectors of the query scene and those of all the other scenes was calculated for each signal. The sum of the ranks of the six signals was calculated, and the five scenes that had the lowest summation were retrieved as similar scenes.

14.3.2 Retrieval Performance Evaluation

To evaluate these methods, we conducted a driving scene retrieval experiment using driving data collected on city roads from 74 drivers (35 males and 39 females). There was about 45 min of recorded driving data per driver, for a total of about 54 h of driving data. The sampling rate of the driving signals was 10 Hz.

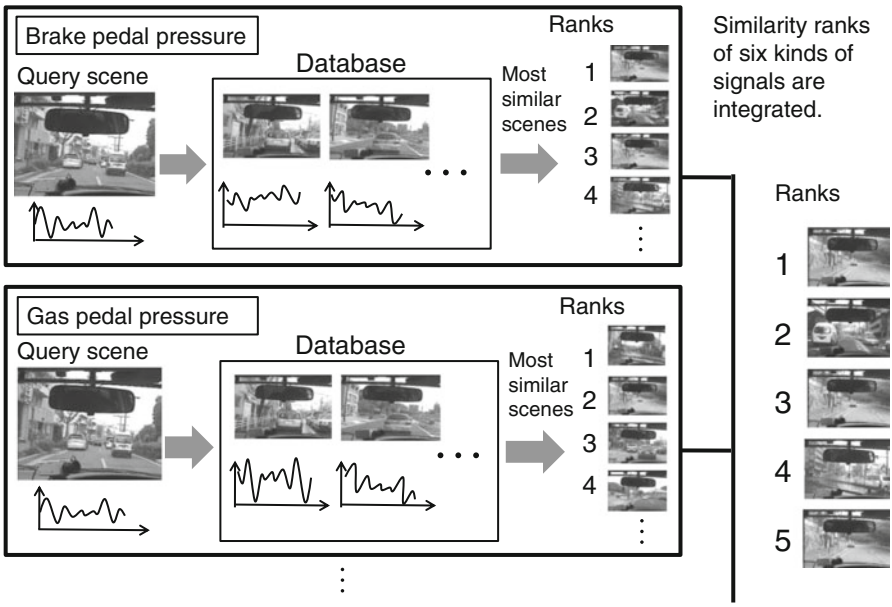


Fig. 14.3 Driving scene retrieval using driving behavior signals (late integration)

14.3.2.1 Experimental Condition

Eight kinds of driving events (stops, starts, right and left turns, right and left lane changes, and going up and down hills) were selected as query scenes, and similar scenes were retrieved using the two techniques described in Sect. 14.3.1. Scenes occurring less than 2 s before or after the query scene, and scenes which had already been retrieved, were excluded from being candidates for retrieval. We chose a total of 80 query scenes, which included about 10 scenes for each type of event.

Retrieval performance was evaluated in terms of retrieval accuracy, i.e., the percentage of correctly retrieved scenes in proportion to the total number of retrieved scenes. Whether or not a scene was correctly retrieved was determined subjectively by human validation.

14.3.2.2 Results

Experimental results are shown in Fig. 14.4. Retrieval accuracy averaging more than 95 % was achieved for driving scenes of stops, starts, and right and left turns, while accuracy was relatively lower for scenes of right and left lane changes, and going up and down hills. Retrieval accuracy of situations involving right turns was higher using the early integration method, but for scenes going down hills, the late integration method was more accurate. On average, the early integration method gave slightly better performance.

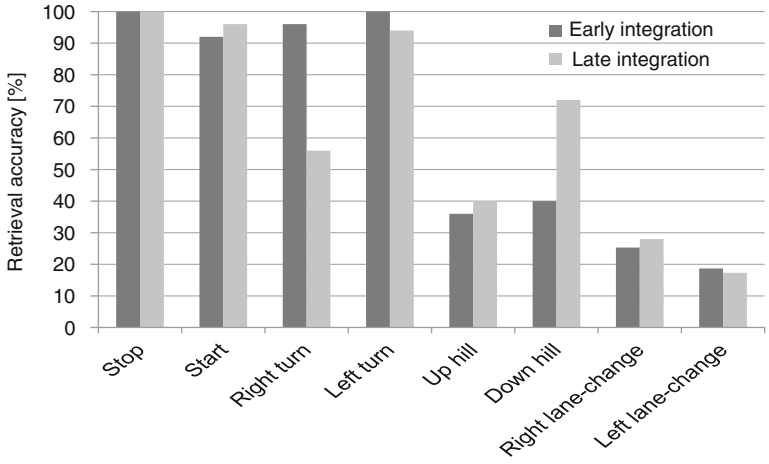


Fig. 14.4 Retrieval accuracy for driving behavior signals

14.4 Driving Scene Retrieval Using Environmental Driving Signals

In contrast to the first study, which employed in-vehicle driving signals, in this section we measured the similarity between scenes by comparing driving environments as detected by laser scanners.

14.4.1 Laser Data Preprocessing

14.4.1.1 Clustering of Laser Data and Tracking of Vehicles

The first step towards automatic scene retrieval was the clustering of discrete laser dots obtained with laser scanners from surrounding driving environments. Each cluster was a set of distance measurements in a plane, grouped closely to each other, and thus probably belonging to a single object. While many approaches have been used to calculate such physical distances [5], we simply used Euclidean distance here. Due to laser dot detection errors, not every cluster actually represented a separate object, i.e., sometimes more than one cluster could belong to a single object. Since all of the laser data were recorded on expressways in this study, in most cases a laser dot must belong to either a vehicle or a roadside barrier, so it was not difficult to integrate some clusters with our prior knowledge of the shapes of these objects [6]. Then, each surrounding vehicle was modeled as a rigid box, characterized by its orientation, position, and velocity. By tracking the vehicles with a Kalman filter, we estimated their dynamic features, even if they were outside the range of the laser scanners.

14.4.1.2 Frame Categorization

A frame categorization method was used to categorize laser-acquired driving frames based on three general features, in order to reduce the number of candidates and facilitate fast retrieval. The scenes were categorized based on road type, congestion level, and the relative positions of surroundings objects. The three features were defined as follows:

- Road type was divided into three classes: left curve, straight line, and right curve. Since two laser scanners were used, one on the front of the vehicle and one on the back bumper, they collected information about road types separately. Their combined data was used to define the road type for each frame of a driving scene, for example, "left curve, straight."
- Road congestion level was divided into two classes: "free flow" and "traffic jam." A Greenshields model [7] was employed to estimate the congestion level for each lane. The road congestion level of a driving frame was designated "traffic jam" if any lane in the frame was estimated as "traffic jam"; otherwise, the frame was designated "free flow."
- Relative positions of surrounding vehicles were classified into 450 situations based on whether there was another vehicle in each of eight surrounding directions and whether there was a roadside barrier on the left or right of the driver's vehicle.

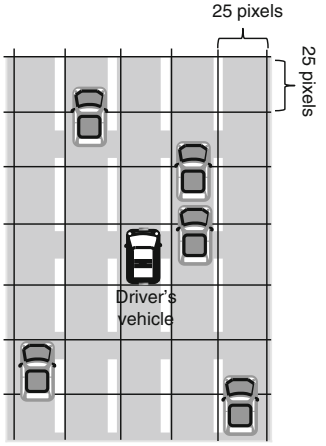
For example, a frame could be represented as "(left curve, straight)," "traffic jam," and "21."

14.4.2 Similarity Measure for Surrounding Environment

Here, we measured the similarity between driving scenes based on the surrounding environment, using three procedures: first, their frame categories (given in Sect. 14.4.1.2) were compared; second, the relative positions of the surrounding vehicles were calculated; and finally, their motion features were compared.

14.4.2.1 Comparison of Frame Categories

In this study, each driving scene consisted of 100 frames (10 s), so each scene could be represented as a vector with 400 dimensions. We then calculated the difference between scenes using Hamming distance to measure how similar the frame categories of two scenes were. Hamming distance between two elements of the vectors took 0 only when the compared features were exactly the same. If the two features



$$G = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 14.5 Example of a frame and its matrix. *Left*: Each cell of the grid is composed of 25 × 25 pixels. The grid is centered on the host vehicle. *Right*: The value of each element of matrix represents the number of vehicles in the corresponding cell

were different, the value was 1. So, if the total Hamming distance was 0, two scenes were identical, and if the total value was 400, they were completely dissimilar. The scenes with a Hamming distance below a threshold of 150 were extracted as candidates for further processing.

14.4.2.2 Comparison of Surrounding Vehicle Positions

The second step was to compare the positions of vehicles in key frames of two scenes. We assumed here that the first frames of scenes were key frames because people generally focus on the first frames of scenes. As shown in Fig. 14.5, a key frame was divided into a grid, and the frame was represented as a matrix G. Each cell of the matrix shows the number of vehicles in the corresponding cell of the grid.

Assume that frames F_1 and F_2 are represented by symbolized matrices G_1 and G_2. To compute the similarity of the two matrices, we first matched all cells in the two frames:

$$\Delta G(F_1, F_2) = \sum_i \sum_j \left| g_{i,j}^{(1)} - g_{i,j}^{(2)} \right|, \tag{14.1}$$

where $g_{i,j}^{(1)}$ and $g_{i,j}^{(2)}$ denote the number of vehicles in cell (i,j) in G_1 and G_2, respectively, and the value of ΔG represents the distance between them. For instance, we can say frames F_1 and F_2 match perfectly if and only if the value of ΔG equals zero. However, this rarely happens because even if two frames are

almost identical, this symbolization method sometimes puts vehicles with the similar positions into different cells. To decrease errors caused by such problems, we also allowed soft matching. We assumed vehicles in two frames matched if there were the same numbers of vehicles in the cells at the same position in two matrices. In addition, we also considered vehicles to match if there were an equal number of vehicles in nearby cells, using a cost function. Thus, the final distance between frames F_1 and F_2 is defined as

$$d(F_1, F_2) = \Delta G^{'}(F_1, F_2) + \frac{k}{K}, \tag{14.2}$$

where $\Delta G'$ is the value of ΔG after soft matching; k is the number of soft matches in $\Delta G'$, and K is an empirically defined normalization factor for the penalty of soft matching.

After that, distance $d(F_1,F_2)$ was used to calculate the similarity between F_1 and F_2:

$$s(F_1, F_2) = \frac{d(F_1, F_2)}{n_1 + n_2}, \tag{14.3}$$

where n_1 and n_2 denote the numbers of vehicles in frames F_1 and F_2, respectively. Frames with a distance below 0.5 from the first frame of a query scene, as well as between their preceding and following frames within 2 s, were selected as key frames for the next step in processing.

14.4.2.3 Comparison of Surrounding Vehicle Motion

If surrounding vehicles have nearly the same positions in the first frames of scenes, as well as similar trajectories and velocities, we believe there is a high probability that these are matching scenes. Also, comparing the motion of surrounding vehicles overcomes problems caused by grid division and achieves a faster search than with frame-to-frame matching between scenes.

Assume that scenes S_1 and S_2 are represented by their vehicle sets (excluding the host vehicle), $V_1 = \left\{ v_1^{(1)}, v_2^{(1)}, \ldots, v_M^{(1)} \right\}$ and $V_2 = \left\{ v_1^{(2)}, v_2^{(2)}, \ldots, v_N^{(2)} \right\}$, where M and N are total numbers of surrounding vehicles observed in S_1 and S_2. At point in time, t, each surrounding vehicle, $v_i^{(1)}$ or $v_j^{(2)}$, is represented as a sequence of vehicle motion feature vectors, consisting of longitudinal position y_i and lateral position x_i with their first-order dynamics Δy_i and Δx_i:

$$(y_i(t), x_i(t), \Delta y_i(t), \Delta x_i(t))^{\mathrm{T}}. \tag{14.4}$$

Dynamic features were calculated by the following equation:

$$\Delta y_i(t) = \frac{\sum_{l=-L}^{L} l \cdot y_i(t+l)}{\sum_{l=-L}^{L} l^2}, \tag{14.5}$$

in which $y_i(t)$ is the ith vehicle's driving signal at point in time t, and L is window size for linear regression. $\Delta x_i(t)$ was calculated in the same way. The distance between vehicles $v_i^{(1)}$ and $v_j^{(2)}$ in two scenes S_1 and S_2, respectively, were calculated as a Mahalanobis distance:

$$d^2\left(v_i^{(1)}, v_j^{(2)}\right) = \left(\mu_{v_i^{(1)}} - \mu_{v_j^{(2)}}\right)^{\mathrm{T}} \Sigma^{-1}_{v_i^{(1)}, v_j^{(2)}} \left(\mu_{v_i^{(1)}} - \mu_{v_j^{(2)}}\right), \tag{14.6}$$

where μ_v represents a four-dimensional vector (including the means of longitudinal position, lateral position, as well as their first-order dynamics) of a vehicle v, and $\sum_{v_i^{(1)}, v_j^{(2)}}$ is a four-by-four covariance matrix of the four features for vehicle $v_i^{(1)}$ and $v_j^{(2)}$. This calculates the distance between a pair of vehicles by comparing the distribution of their four-dimensional features. Based on our preliminary experiment, a pair of vehicles with a Mahalanobis distance below a threshold of 15.0 was believed to be similar to each other.

To acquire a vehicle-to-vehicle match, we calculated $d(v_i, v_j')$ for all i and j between scenes and selected them from smallest to largest. We considered scenes to be similar to each other if there were enough similar vehicles in both scenes. Similarity p between S_1 and S_2 was defined as the summation of the weights of similar vehicles divided by the summation of the weights of all the vehicles in the scenes:

$$p(S_1, S_2) = \frac{\sum_{i \in X_1} \sum_{t \in T_i^{(1)}} w_t^{(i)} + \sum_{i \in X_2} \sum_{t \in T_i^{(2)}} w_t^{(i)}}{\sum_{i \in Y_1} \sum_{t \in T_i^{(1)}} w_t^{(i)} + \sum_{i \in Y_2} \sum_{t \in T_i^{(2)}} w_t^{(i)}}, \tag{14.7}$$

where X_1 and X_2 denote the sets of similar vehicles, and Y_1 and Y_2 denote the sets of all vehicles in S_1 and S_2, respectively. $w_t^{i)}$ denotes the weight of vehicle v_i at time t. $T_i^{(1)}$ and $T_i^{(2)}$ are the sets of frame numbers where $v_i^{(1)}$ or $v_i^{(2)}$ was observed in S_1 or S_2, respectively. Here, "weight" means the importance of a surrounding vehicle, which was represented as a value of a modified Gaussian distribution as illustrated in Fig. 14.6. The reason we used a modified Gaussian distribution which was stretched towards the front value as a similarity metric is that, generally, a driver is more aware of nearby leading vehicles while driving. For example, the

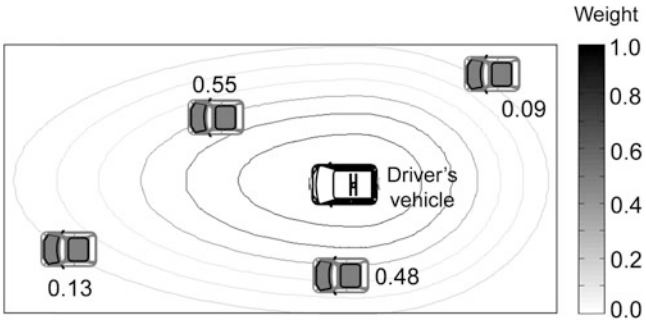


Fig. 14.6 A modified two-dimensional Gaussian distribution, centered on the driver's vehicle, where surrounding vehicles with higher values denote greater importance

surrounding vehicles in front of a driver's vehicle are more important than those on either side of or behind the driver's vehicle. It can be inferred that a pair of similar vehicles near the driver's vehicle makes scenes more similar than pairs located farther away.

14.4.3 Retrieval Performance Evaluation

The proposed driving scene retrieval system was evaluated using database-containing expressway scenes from 57 drivers (28 males and 29 females) recorded with the instrumented vehicle shown in Fig. 14.1. The database contained approximately 140,000 driving frames. All of the driving data were sampled at 10 Hz. We compared retrieval accuracy and speed for different types of scenes under various retrieval conditions, by using subjective scores and by measuring retrieval speed in CPU time. Here, "retrieval conditions" mean some combinations of the similarity measures presented in Sect. 14.4.2:

(a) Based on frame category
(b) Based on surrounding vehicle position
(c) Based on surrounding vehicle motion

The combinations are represented as a, c, $a + c$, $b + c$, and $a + b + c$. We did not use b or $a + b$, since b only considered the first frame of a scene and would not be accurate if used alone.

The experiment was conducted as follows:

- Five driving scenes each, for straight road, curve, traffic jam, and lane change, were randomly selected as queries.
- For each query scene, we evaluated retrieval accuracy and retrieval speed for each retrieval condition. For each condition, the top five similar scenes were retrieved, and they were used for the evaluation.

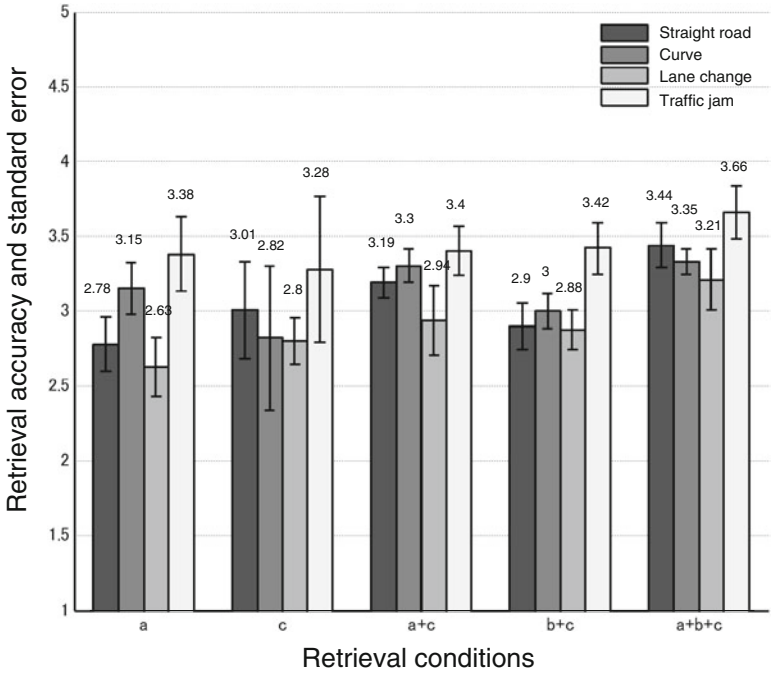


Fig. 14.7 Comparison of retrieval accuracy

14.4.3.1 Comparison of Retrieval Accuracy Using Subjective Scores

In this comparison, the subjective scores of five volunteers were used to judge which retrieval condition, or combination of retrieval conditions, was able to select scenes with the highest similarity to a query scene for a given driving situation. Each volunteer gave scores, from 1 (lowest) to 5 (highest), to the top five retrieved scenes for each query under each retrieval condition. Scenes with a score of 3 or higher were considered to be similar. A score of 5 indicated perfect similarity, while a score of 1 indicated complete dissimilarity. The retrieval accuracy of a given scene under a given retrieval condition was estimated as the average of the scores from the five volunteers.

The experimental results are shown in Fig. 14.7, which indicate that condition $a + b + c$ demonstrated much higher accuracy than the other conditions, in various driving situations.

14.4.3.2 Comparison of Retrieval Speed Using CPU Time

In order to compare processing speed, the proposed driving scene retrieval system was installed on a Core i5 CPU 650@3.20 GHz PC using the Windows 7 operating system. The CPU time for each query process was recorded for each retrieval condition.

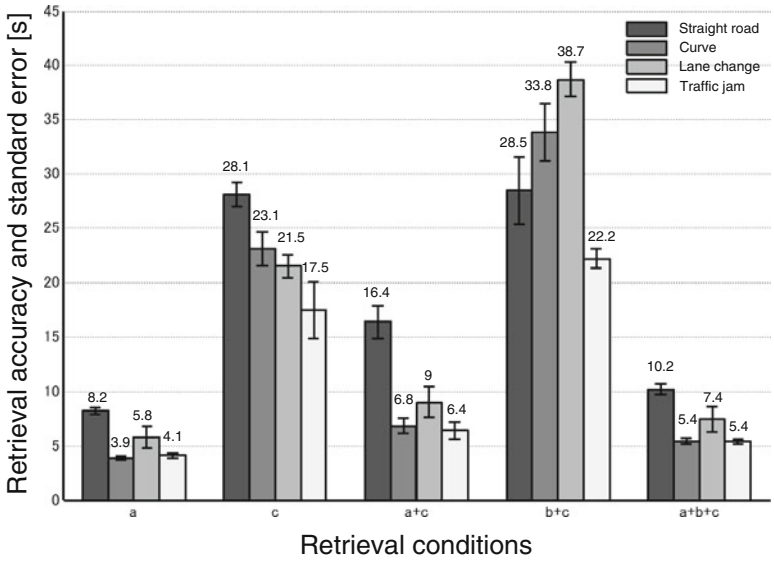


Fig. 14.8 Comparison of retrieval speed

The average retrieval time for top five driving scenes was calculated. This was considered to represent system speed performance under a given retrieval condition for each scene. Figure 14.8 shows the average retrieval time taken to retrieve scenes from the 140,000-frame database. On average, retrieval condition a took the least time, and condition $a + b + c$ was the next fastest.

14.5 Conclusions

In this chapter, we developed two systems for retrieving recorded driving scenes based on measuring the similarity of driving behavior and environmental driving signals. In the first study, similar scenes were retrieved using driving behavior signals, and they were integrated using two methods, early and late integration. Experimental results showed that an average of more than 95 % retrieval accuracy was achieved for driving scenes of stops, starts, and right and left turns. In most situations, the early integration method achieved better performance than the late integration method. In the second study, we used environmental driving signals with the idea that similar driving scenes could be retrieved by measuring similarity in surrounding environments. Experimental results showed that the integrated use of information from surrounding vehicles and road conditions achieved higher retrieval accuracy than the use of either type of information alone.

Currently, we are working to integrate these two systems, to see if retrieval accuracy can be further improved.

Acknowledgement This work was partially supported by the Strategic Information and Communications R & D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan under No. 082006002, by Grant-in-Aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS) under No. 24500200, and by the Core Research of Evolutional Science and Technology (CREST) of the Japan Science and Technology Agency (JST).

References

1. S.Y. Cheng, S. Park, M.M. Trivedi, Multispectral and multi-perspective video arrays for driver body tracking and activity analysis. Comput. Vis. Image Understand. **106**, 245–247 (2007)
2. D. Mitrovic, Reliable method for driving events recognition. IEEE Trans. Intell. Transp. Syst. **6** (2), 198–205 (2005)
3. N. Oliver, A. Pentland, Graphical models for driver behavior recognition in a SmartCar, in *Proceedings of the IEEE Intelligent Vehicles Symposium,* pp. 7–12, 2000
4. M. Naito, A. Ozaki, C. Miyajima, N. Kitaoka, R. Terashima, K. Takeda, A browsing and retrieval system for driving data, in *Proceedings of the IEEE Intelligent Vehicles Symposium,* pp. 1159–1165, June 2010
5. J.Z. Wang, J. Li, G. Wiederhold, SIMPLIcity: semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. **23**(9), 947–963 (2001)
6. N. Kaempchen, Feature-level fusion of laser scanner and video data scanner and video for advanced driver assistance systems, Ph.D. dissertation, University of Ulm, Germany, 2007
7. H. Rakha, B. Crowther, Comparison of Greenshields, pipes, and Van aerde car-following and traffic stream models. J. Transp. Res. Board **1802**, 248–262 (2007)
8. K. Takeda, J. Hansen, P. Boyraz, L. Malta, C. Miyajima, H. Abut, International large-scale vehicle corpora for research on driver behavior on the road. IEEE Trans. Intell. Transp. Syst. **12**, 1609–1623 (2011)