Dashcam Video Curation for Generating Memorial Movies on Tourism using Multiple Measures of "Tourist Spot Likeness"*

Masaki Kawanaka[1], Yuki Matsuda[1,2][0000−0002−3135−4915], Hirohiko Suwa[1,2][0000−0002−8519−3352], and Keiichi Yasumoto[1,2][0000−0003−1579−3237]

[1] Nara Institute of Science and Technology, Nara 630-0101, Japan
{kawanaka.masaki.kj1, yukimat, h-suwa, yasumoto}@is.naist.jp
[2] RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan

Abstract. Recently, a method for the automatic generation of driving sightseeing memorial route movies has been proposed by performing curation on videos captured by a dashcam. Existing methods curate videos by scoring video frames with a single specific measure, however, it is hard to generate the appropriate memorial route movies due to the differences in important perspectives for different users. To solve this problem, we propose the method that represents "tourist spot likeness" with a combination of common measures derived by procedures used in the field of Affective Engineering and curates dashcam videos based on multiple measures. In this paper, we surveyed these representative words to identify the measures that compose the measure of "tourist spot likeness" and construct a model to estimate the score of each measure using dashcam video data as input. As the results of the experiment by using dashcam videos taken in Okinawa, we have derived six measures for curating the dashcam videos and confirmed the proposed models, which estimate each measure scored with a 7-level Likert scale achieved mean MAE of 0.64 (best: 0.52, worst: 0.73).

Keywords: Affective Engineering, Civic and Urban Computing, Human-Centered Artificial Intelligence, Location-based Services, Smart Tourism, Video Curation, Dashcam

1 Introduction

In recent years, an increasing number of travelers have been posting memorial videos created from photos and videos of their travels on social media. Most of the memorial movie generation support systems, such as RealTimes [10], compile photos of multiple tourist spots. Demand for such systems that automatically create memorial videos is growing because they are easy to use, even for users without video editing skills. However, few services provide memorial movie generation support systems focusing on tourist routes, especially while driving.

* This study was supported in part by JST PRESTO under Grant No. JPMJPR2039.

Since the trip between tourist spots constitutes a large percentage of sightseeing, many important scenes exist. It can be considered that the memorial tourist movie generation support system focusing on tourist routes helps to share tourist flow more intuitively than the system curates photos taken at multiple tourist spots. In addition to public transportation such as buses and trains, walking and driving are also considered as ways of tourist transportation. In sightseeing using public transportation such as buses and trains, it is difficult to capture tourist routes, but in the case of cars, it is possible to capture tourist routes using a dashcam. As the performance of dashcams improves, they are not only used as evidence recorders in the event of accidents. Still, they are also utilized in tourism support [9], driving skill improvement systems [12], and on-street parking detection [7].

Katayama *et al.* [6] have developed a video curation algorithm to preserve memorable scenes from a dashcam in a memorial video. The algorithm removes redundant parts, such as irreplaceable road trips, stops at traffic lights, and traffic jams, and extracts memorable scenes that are the highlights of the video. To calculate the importance of the frame, they propose a measure named "Okinawa likeness", which is the scale of how well the video frame matches the impression of Okinawa (a prefecture in Japan). Then, the "Okinawa likeness" of dashcam videos taken during sightseeing tours in Okinawa is collected and measured using crowdsourcing, and a machine learning model is constructed to predict "Okinawa likeness" for memorial video curation. However, ambiguous expressions such as "Okinawa likeness" are perceived differently by different people [8], making it difficult to generate appropriate memorial route movies based on the various "Okinawa likeness" assumed by each user, and also difficult to apply the model to other touristic areas.

This paper aims to improve the quality of curation of memorial route movies by expressing "`tourist spot likeness`" using a combination of general-purpose measures that are less prone to differences in interpretation among people, instead of measures that are prone to differences in interpretation used in previous studies. To achieve this, it is necessary to clarify what combination of impressions people get from tourist spots can express "`tourist spot likeness`". In this paper, we use dashcam videos taken in Okinawa prefecture, one of Japan's most touristic southern islands. To derive a measure to be used for curating the memorial route movie, we extract clusters of impression words that compose "`tourist spot likeness`" and select representative words based on procedures used in the field of Affective Engineering. Then, we use these representative words to identify measures of "`tourist spot likeness`", and construct a CNN model that estimates scores for each measure using dashcam videos. As a result, we identified six measures necessary for dashcam video curation, and confirmed proposed models, which estimate each measure scored with a 7-level Likert scale achieved mean MAE of 0.64 (best: 0.52, worst: 0.73).

This paper is organized as follows: In Section 2, we introduce related studies and summarize the position of this research. Next, in Section 3, we describe a memorial video curation method using drive recorder images. In Section 4, we

describe a method for identifying a measure of "`tourist spot likeness`" using the results of extracting representative words that compose the measure, and in Section 5, we construct and evaluate a CNN model for estimating the score of each measure. Finally, Section 6 concludes this paper.

2 Related Research

2.1 Video curation for route guidance

A system that provides route guidance must correctly communicate walking and driving routes. Although maps are a simple route guidance system, some users may have difficulty understanding a route with only a map. Therefore, methods have been proposed to facilitate understanding of routes by utilizing curated videos for route guidance [5]. In these methods, the playback speed is increased for parts that are not necessary for route guidance, such as going straight, and curated videos are played at normal speed for parts that are important for route guidance, such as turning right or left. In a study for walking route guidance[5], the first-person video is used to detect landscape transitions based on histogram differences, and curation is performed using a variable frame rate method. For driving route guidance, we use video from a drive recorder to compute an optical flow based on the LucasKanade algorithm to detect left-right turns and curate videos.

2.2 Video curation for memorial video creation

In recent years, an increasing number of travelers have been posting memorial videos created from photos and videos of their travels on social media. However, the creation of memorial videos using video editing software is difficult for users who do not have video editing skills. For this reason, systems have been proposed to automatically create memorial videos easily for people without video editing skills [10]. However, memorial videos generated by conventional systems lack scenes during travel, a major element of tourism, and fail to reflect the impressive scenes that occur during travel. Therefore, an automatic curation method for memorial videos of the entire tourist tour, including moving around, has been proposed using dashcam video.

In the method of Katayama *et al.* [6], dashcam videos of tourist routes are segmented into 3-second segments, and the importance of each segment is estimated to evaluate whether it is necessary as a tourist memorial video, and memorial video curation is performed. As an importance estimation model, three image frames are extracted from each 3-second segment video, and an estimation model of "Okinawa likeness" was constructed using category occupancy calculated by DeepLabv3 [2] trained with CityScapes [3] and BDD100k [14], and landmark information obtained by YOLOv3 [11] as features. Crowdsourcing was used to evaluate the "Okinawa likeness" of the segmented videos, which were then used as the correct labels for the scores. However, since the vague expression

"Okinawa likeness" is perceived differently by different people, it may not be possible to generate memorial route movies based on the "Okinawa likeness" assumed by each user.

2.3 Position of this research

In this study, we aim to improve the quality of memorial video curation by expressing "`tourist spot likeness`" using a combination of measures that are less likely to cause differences in interpretation among people, instead of a measure that is more likely to cause differences in interpretation, based on the video curation method using the importance score proposed in Section 2.2. We extract clusters of impression words that compose "`tourist spot likeness`" and select representative words based on procedures used in the field of Affective Engineering [13], to derive a measure for use in curating memorial route movies using dashcam videos of tourist tours in Okinawa Prefecture.

In this paper, we use these representative words and factor analysis to identify the measures that compose "`tourist spot likeness`". Then, we construct a CNN model to estimate the scores of each measure, and evaluate and discuss whether the scores of each measure can be estimated or not.

3 Proposed System

The overview of the proposed method is shown in Fig. 1. First, the dashcam video is split into 3-second segments, and the first, middle, and last frames are extracted. When dividing the dashcam video into segmented videos, we divided the video every 3 seconds, taking into account the characteristic that most people's visual information enters a stable period in 3 seconds [1]. Three seconds is the approximate time it takes for an object 50m ahead to frame out from the view of a car traveling at 60km per hour, which is enough time to confirm changes in the scenery outside the car. Next, the estimated scores of each measure are obtained from machine learning models that estimate the measures that compose "`tourist spot likeness`". Finally, segments are selected based on the score of each measure, and a memorial tourist movie is generated.

The importance score used for selecting a segment is calculated from a combination of several measures, as shown in Fig. 2. In the proposed method, the "`tourist spot likeness`" is defined as follows, using the measures that compose "`tourist spot likeness`" and the weight parameter.

$$S = \frac{1}{L}\sum_{l=1}^{L} s_l w_l \quad , \tag{1}$$

where S denotes the "`tourist spot likeness`", L denotes the total number of measures, w denotes the weight parameter of measure that composes "`tourist spot likeness`". The proposed method can freely change the weight parameter w, and can generate a memorial route movie based on each user's idea of "`tourist spot likeness`".

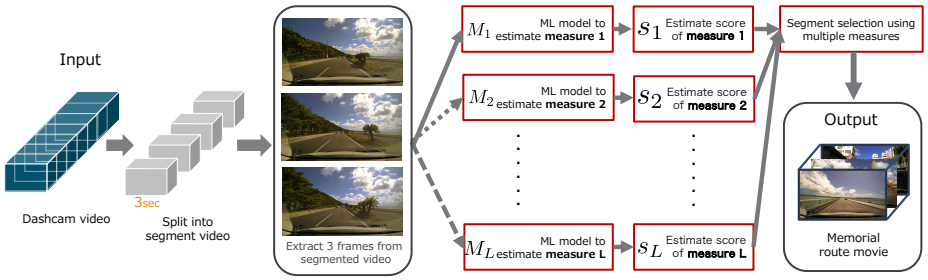

Fig. 1. Overview of proposed method

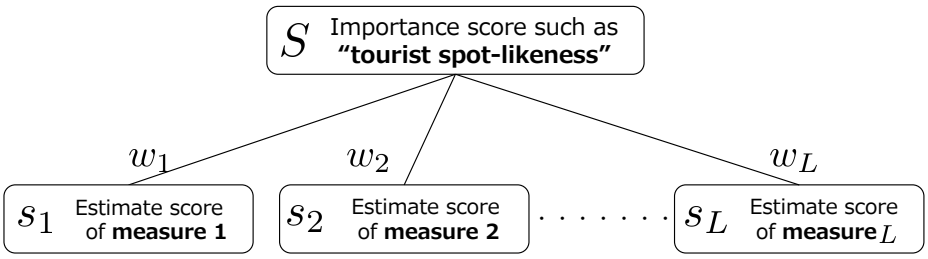

Fig. 2. Measures that compose "`tourist spot likeness`"

4 Selection of measures of "tourist spot likeness"

To realize the proposed method, it is necessary to clarify what kinds of combination of impressions that people receive from tourist spots can effectively express "`tourist spot likeness`". This section describes the method to derive "`tourist spot likeness`" measure.

Hereinafter, we describe the proposed method with an example of dashcam videos taken in Okinawa prefecture, one of Japan's most touristic southern islands.

4.1 Overview of measure derivation method

To derive the measures of "`tourist spot likeness`" to be used for the curation of memorial route movies, we employ the following procedure inspired by the method in the field of Affective Engineering: (1) Collecting the impression words when watching the dashcam video, (2) Conduct a psychological experiment to calculate the distance in semantic space between impression words, extract clusters of impression words expressing "`tourist spot likeness`" by hierarchical cluster analysis based on the obtained distance, and select representative words, (3) Using the representative words obtained and factor analysis, a measure of "`tourist spot likeness`" is derived.

Table 1. Representative words that constitute "tourist spot likeness" for Okinawa prefecture in Japan

Cluster 1	Cluster 2	Cluster 2	Cluster 4	Cluster 5	Cluster 6
radiant (晴れやかな)	comfortable (気持ちいい)	relax (のんびりした)	seaside (海沿いの)	suburban (郊外の)	peaceful (平穏な)
brightly (明るい)	pleasant (心地良い)	loosely (ゆったりとした)	coastal (海岸沿いの)	quiet (閑静な)	composed (落ち着いた)
fresh (清々しい)		calm (穏やかな)	refined (優雅な)	routine (日常の)	silent (静かな)

Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12
spacious (開放的な)	warm (暖かい)	dazzling (眩しい)	expressway (高速道路の)	exotic (異国情緒な)	lively (賑やかな)
open (開けた)	animated (生き生きした)	summer (常夏の)	night (夜の)	foreign (異国の)	hot (暑い)
vast (広大な)	tropical (南国な)	new (新しい)	roadside (道路沿いの)	Asian (アジア風の)	

Cluster 13	Cluster 14	Cluster 15	Cluster 16
messy (ごみごみした)	trafficy (渋滞の)	urban (都会の)	dark (暗い)
congestion (渋滞した)	crowded (雑踏の)	downtown (街中の)	narrow (狭い)
stuffy (息苦しい)	cramped (窮屈な)		darkness (暗がりの)

First, we have conducted procedures (1) and (2) using dashcam videos of Okinawa, and the representative words of each cluster obtained by cluster extraction are shown in Table 1.

4.2 Derivation of measures of "tourist spot likeness"

In order to derive a measure of "`tourist spot likeness`", we used Yahoo! Crowdsourcing[3] to conduct a task in which respondents were asked to rate, on a 7-point scale (1. Strongly agree, 2. agree, 3. Slightly agree, 4. Neither agree nor disagree, 5. Slightly disagree, 6. disagree, 7. Strongly disagree), how they felt about the words shown in Table 1 when watching a dashcam video of Okinawa. The videos used in this experiment were fifty 3-second segment videos generated from dashcam videos taken in Okinawa. In the segment videos, we used dashcam videos of various scenes, as shown in Fig. 3, so that various impression words could be collected. For each video, 10 participants rated how they felt about the three representative words that were presented when they watched the video. If a person did not correctly answer a check question, all of his or her data were discarded.

[3] https://crowdsourcing.yahoo.co.jp/

Fig. 3. Segment video created from drive recorder video taken in Okinawa

A factor analysis using the maximum likelihood method and Promax Rotation was then performed on the means of each evaluation score for each of the 10 videos obtained by crowdsourcing. The number of factors was determined to be 7, based on the criterion that the eigenvalue must be greater than 1. However, as the number of measures increases, the usability of the curation system may decrease because users are required to select which measure is more important when using the system. Therefore, the number of factors in this experiment was set to 6.

The top three absolute values of factor loadings (FL) for representative words showing positive and negative correlations for each factor obtained by factor analysis are shown in Table 2. Here, we define and name the measure of "`tourist spot likeness`" for each factor. Factor 1 is positively correlated with words that indicate less crowdedness, such as "quiet" and "silent", and negatively correlated with words that indicate more crowdedness, such as "congestion" and "messy", thus Factor 1 is named as "`urbanness`". Factor 2 has a positive correlation with words that describe the morning and afternoon hours, such as "brightly" and "radiant", and a negative correlation with words that describe the evening hours, such as "darkness" and "night", thus Factor 2 is named as "`brightness`". Factor 3 is named as "`exoticness`" because of its strong correlation with words that indicate exoticism, such as "exotic" and "tropical", and Factor 4 is named as "`scenicness`" because of its strong correlation with words that indicate scenery, such as "seaside" and "coastal". Factor 5 is positively correlated with words that indicate more exhilaration, such as "expressway", and negatively correlated with words that indicate less exhilaration, such as "downtown" and "relax", thus Factor 5 is named as "`liveliness`". Factor 6 has a positive correlation with words that describe the open view, such as "open" and "vast", and a negative correlation with words that describe the less open view, such as "narrow" and "cramped", thus Factor 6 is named as "`openness`". The results of the factor analysis indicate that the six measures of "`tourist spot likeness`" are "`urbanness`", "`brightness`", "`exoticness`", "`scenicness`", "`liveliness`", and "`openness`".

Table 2. Top three absolute magnitudes of factor loadings (FL) for representative words showing positive and negative correlations for each factor, and measure names

	Factor 1	FL	Factor 2	FL	Factor 3	FL
positive correlation	quiet (閑静な)	0.98	brightly (明るい)	0.87	exotic (異国情緒な)	0.91
	silent (静かな)	0.91	radiant (晴れやかな)	0.82	foreign (異国の)	0.88
	relax (のんびりした)	0.87	warm (温かい)	0.51	tropical (南国な)	0.974
negative correlation	congestion (渋滞した)	-0.99	darkness (暗がりの)	-0.97	expressway (高速道路の)	-0.43
	trafficy (渋滞の)	-0.91	dark (暗い)	-0.74	suburban (郊外の)	-0.27
	messy (ごみごみした)	-0.88	night (夜の)	-0.72	congestion (渋滞した)	-0.14
	⇓ “urbanness”		⇓ “brightness”		⇓ “exoticness”	

	Factor 4	FL	Factor 5	FL	Factor 6	FL
positive correlation	seaside (海沿いの)	0.92	expressway (高速道路の)	0.86	open (開けた)	0.47
	coastal (海岸沿いの)	0.91	new (新しい)	0.50	roadside (道路沿いの)	0.43
	vast (広大な)	0.35	dazzling (眩しい)	0.37	vast (広大な)	0.42
negative correlation	routine (日常の)	-0.35	routine (日常の)	-0.42	narrow (狭い)	-0.81
	lively (賑やかな)	-0.24	downtown (街中の)	-0.36	cramped (窮屈な)	-0.29
	narrow (狭い)	-0.16	relax (のんびりした)	-0.33	Asia-like (アジア風の)	-0.26
	⇓ “scenicness”		⇓ “liveliness”		⇓ “openness”	

5 Building of “tourist spot likeness” estimation model

In this section, we build and evaluate a model to estimate scores using dashcam video data as input for the six measures of “`tourist spot likeness`” derived in Section 4.

5.1 Dataset

In this experiment, we use dashcam videos taken over four days from July 11 to 14, 2020 in Okinawa. When the dashcam video was divided into 3-second segmented videos, 459 videos were generated on the first day, 831 videos on the second day, 1,244 videos on the third day, and 569 videos on the fourth day. In total, 3,103 videos are included in the dataset. We collect the ground-truth

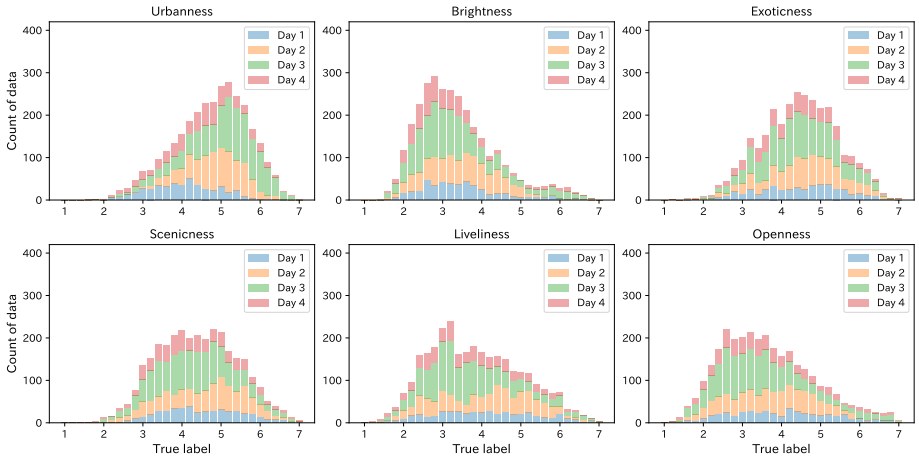

Fig. 4. Relationship between the number of true labels and the number of data in each session

labels of how people perceive the measure of "`tourist spot likeness`" when watching the videos, by using Yahoo! Crowdsourcing. Each video is annotated by five people, and the ground-truth label is derived by averaging them. If a person did not correctly answer a dummy question for detecting careless responses, all of their data were discarded and were not included in the following analysis. To inform the meaning of each measure, we provided the following explanations to crowd workers in advance. The "`urbanness`" measure is closer to 1, the more urban it is, and the closer it is to 7, the more rural it is. The "`brightness`" measure is closer to 1 for brighter, and closer to 7 for darker. The "`exoticness`" measure is closer to 1, indicating a sense of exoticism, and closer to 7, indicating a lack of a sense of exoticism. The "`scenicness`" measure is closer to 1, indicating beautiful scenery, and closer to 7, indicating less beautiful scenery. The "`liveliness`" measure is closer to 1, indicating exhilaration, and closer to 7, indicating convergence. The "`openness`" measure is closer to 1, indicating a sense of openness, and closer to 7, indicating a sense of tightness.

5.2 Experimental conditions

In this experiment, we use ResNet [4], which has been shown to be effective in the field of image recognition, to estimate the scores of the measures that constitute "`tourist spot likeness`". The model is a trained model of ResNet50 provided by PyTorch. As shown in Fig. 1, the dashcam video is divided into 3-second segment videos, and three images extracted from the 3-segment videos are used as input. The three images used as input are the first, middle, and last frames of the 3-second segment video. When inputting the images into the ResNet, the three images were concatenated in the dimensionality direction where RGB is stored. As output, the number of output classes was changed to 1 to obtain a score

Table 3. Results of the evaluation of six measure score estimation models（MAE）

urbanness	brightness	exoticness	scenicness	liveliness	openness	Mean
0.73	0.52	0.72	0.65	0.59	0.60	0.64

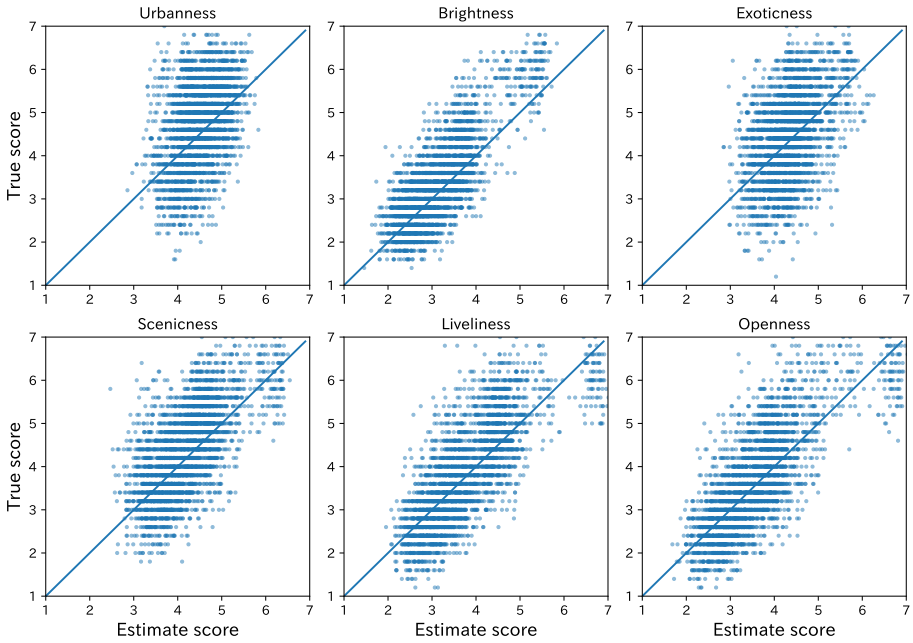

Fig. 5. Relationship between the true values and the estimated results of the six-measure score.

for each measure and treated as a regression model. To evaluate generalization performance, leave-one-day-out cross-validation was conducted with one day's worth of data as the unit. The relationship between the number of true labels and the number of data in each session is shown in Fig. 4. Mean Absolute Error (MAE) was used as the loss function during training, and Stochastic Gradient Descent (SGD) was used as the optimization method with a learning rate of 0.01. The number of mini-batches was 32 and the number of epochs was 20.

5.3 Experimental results

The results of the MAE evaluation are shown in Table 3, and the results of the estimated measure scores using the ResNet are shown in Fig. 5. Table 3 shows that the mean of MAE for all measures is about 0.64, which indicates that the results are relatively good. Furthermore, Fig. 5 shows that the estimated scores for

"`brightness`", "`scenicness`", "`liveliness`", and "`openness`" increased as the score of the true label increased, indicating that the results captured the trend. On the other hand, for "`urbanness`" and "`exoticness`", the estimated scores were concentrated around the center (3–5), indicating that the estimates were not correct. One possible cause of this result is the influence of data bias caused by the division of the test data when conducting the leave-one-day-out cross-validation. As shown in Fig. 4, the data around the true label 6 for "`urbanness`" are more abundant in Day 3 than in the other data, and the model using Day 1, 2, and 4 as training data is not able to correctly learn the data around the true label 6, and thus is considered to be unable to estimate it.

6 Conclusion

The purpose of this study is to improve the quality of memorial video curation by expressing "`tourist spot likeness`" using a combination of general-purpose measures that are less likely to be interpreted differently by different people instead of the measures that have been used in existing studies and that are prone to differences in interpretation. In this paper, we derived six measures of "`tourist spot likeness`" and constructed a CNN model to estimate each measure of dashcam data taken in Okinawa prefecture. The experimental results show that the mean MAE of all the measures is about 0.64, which is relatively good. On the other hand, due to data imbalance problems, it became clear that we could not construct a model that shows sufficient performance for some measures.

References

1. Brady, T.F., Konkle, T., Alvarez, G.A., Oliva, A.: Visual long-term memory has a massive storage capacity for object details. Proceedings of the National Academy of Sciences **105**(38), 14325–14329 (2008). https://doi.org/10.1073/pnas.0803390105
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv preprint arXiv:1706.05587 pp. 1–14 (2017). https://doi.org/10.48550/arXiv.1706.05587
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223. CVPR '16 (2016). https://doi.org/10.1109/CVPR.2016.350
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. CVPR '16 (2016). https://doi.org/10.1109/CVPR.2016.90
5. Kanaya, Y., Kawanaka, S., Suwa, H., Arakawa, Y., Yasumoto, K.: Automatic Route Video Summarization based on Image Analysis for Intuitive Touristic Experience. Sensors and Materials **32**(2), 599–610 (2020). https://doi.org/10.18494/SAM.2020.2616

6. Katayama, Y., Suwa, H., Yasumoto, K.: dash-cum: Dashcam Video Curation for Memorial Movie Generation. In: The 27th Symposium on Information Systems for Society. ISS '21 (2021), (in Japanese)
7. Matsuda, A., Matsui, T., Matsuda, Y., Suwa, H., Yasumoto, K.: A Method for Detecting Street Parking Using Dashboard Camera Videos. Sensors and Materials **33**(1), 17–34 (2021). https://doi.org/10.18494/SAM.2021.2998
8. Matsuda, Y.: IoPT: A Concept of Internet of Perception-aware Things. In: The 12th International Conference on the Internet of Things. pp. 201–204. IoT '22 (2022). https://doi.org/10.1145/3567445.3571108
9. Morishita, S., Maenaka, S., Nagata, D., Tamai, M., Yasumoto, K., Fukukura, T., Sato, K.: SakuraSensor: Quasi-Realtime Cherry-Lined Roads Detection through Participatory Video Sensing by Cars. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 695–705. UbiComp '15 (2015). https://doi.org/10.1145/2750858.2804273
10. RealNetworks: RealTimes. https://jp.real.com/realtimes/, (Accessed on Dec. 28, 2022)
11. Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767 pp. 1–6 (2018). https://doi.org/10.48550/arXiv.1804.02767
12. Takenaka, K., Bando, T., Nagasaka, S., Taniguchi, T.: Drive Video Summarization Based on Double Articulation Structure of Driving Behavior. In: Proceedings of the 20th ACM International Conference on Multimedia. pp. 1169–1172. MM '12 (2012). https://doi.org/10.1145/2393347.2396410
13. Tobitani, K., Matsumoto, T., Tani, Y., Nagata, N.: Modeling the Relation between Skin Attractiveness and Physical Characteristics. In: Proceedings of the 2018 International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia. pp. 30–35. MMArt&ACM '18 (2018). https://doi.org/10.1145/3209693.3209699
14. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2633–2642. CVPR '20 (2020)