特集 Prediction of Driver Operations inside Vehicles *

伊藤隆文　　金出武雄
Takafumi ITO　　Takeo KANADE

Recently, developments of various intelligent vehicles have been performed by installing sophisticated systems with the aim of safety and comfort. In order to realize more sophisticated systems harmonized with drivers, it will be important for the systems to recognize and adapt to the driver's situation or operation. In this paper, we propose a new method for predicting typical driver operations, which are performed by vehicle drivers, such as "pushing navigation buttons", "adjusting the rear-view mirror", or "opening the console box", before the fingers of the drivers actually reach the target position. The prediction method used a camera to capture images of anatomical landmarks (shoulders, elbows, and wrists) when they moved over time. The difference in the configurations of various operations was modeled using a combination of clustering and discriminant analysis. The proposed method was applied to predict the nine most frequently executed operations inside a vehicle, running at over 150 frames per second. The proposed system achieved an average prediction accuracy of 90% with five subjects in a driving simulator.

Key words: Prediction, Pattern recognition, Operation, HMI

1. INTRODUCTION

The primary goal of our researches is to make the experience of driving safer and more comfortable. Then it is important that vehicle anticipates driver behaviors and provides suitable assists and timely information. To enable the smart assistance systems, we require algorithms to predict the behavior of drivers in advance and react preemptively. In such systems, it is imperative that a high level of accuracy is achieved so drivers are able to rely on them with a high degree of confidence.

When operating a vehicle, the driver's primary tasks include turning at intersections, stopping at stop signs, and changing lanes. Secondary tasks not directly related to driving include operating the air conditioner, adjusting the seat, or drinking a beverage. These tasks divert a driver's attention away from primary driving tasks and negatively affect response times. It is therefore necessary to predict such operations to warn drivers of danger earlier and to assist drivers in operating the vehicle more safely. Predictive systems can also lead to improvement in the usability of equipment inside the vehicle.

There are two primary challenges in predicting operations for driver assistance. First, in order for assistance systems to be useful, they must be able to disambiguate operations quickly. The typical duration of an operation inside a vehicle is on the order of a second. The system must be able to operate in a fraction of that time to be able to react usefully. Second, most operations performed by a driver are similar, at least at the beginning of their execution. For example, reaching for the glove compartment and using the center panel both involve the left hand[1] moving in approximately the same direction − at least initially. The challenge is to be able to predict the operation early in its execution while the behaviors are usually still not significantly distinct.

Our goal is to predict which operation a driver is executing from six joint positions (shoulders, elbows and wrists). We are interested in accurate prediction as early as possible in the evolution of an operation. A camera on top of the windshield of the vehicle captures video at 60 frames per second at 640 × 480 − shown in **Fig. 1**(a). The eventual destinations of the left hand in nine operations inside a vehicle are shown in **Fig. 1**(b). The time when the driver moves his or her hand away from the neutral position on the steering wheel is defined as the start time of the operation, and the time when the driver touched the target equipment is the end of the operation, or to be exact, the end of the movement for the operation. The duration of the operation is the period between the two events defined.

In this paper, we use training data to learn models that discriminate each object class and use these models for operation prediction. During training we cluster the training corpus, and we apply multiple discriminant analysis for each cluster in the training set. During execution, we find the probability of association of the current configuration to each cluster and compute an "expectation" of the final operation.

[1]Depending on driving conventions

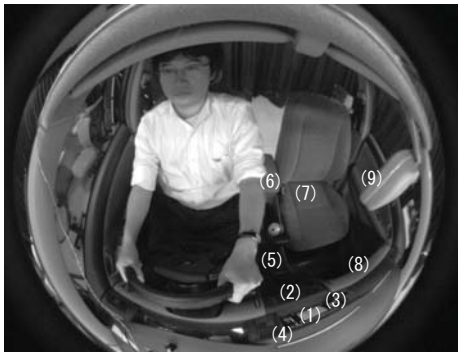
* Reprinted with permission from 8th IEEE Int' l Conference on Automatic Face and Gesture Recognition

The proposed method was applied to predict nine frequently performed operations inside a vehicle. On a standard desktop computer it runs at an average rate of 172 frames per second. For five subjects, it achieves an average prediction accuracy of 90% with a false positive rate of 1.4% after half the operation duration for five subjects. The average duration of an operation was 1.2 seconds and this accuracy was achieved at an average time of 0.52 seconds. The algorithm was compared in performance with HMM-based methods, achieving significantly higher prediction accuracies earlier in the operation.

2. RELATED WORK

The literature on analyzing human behavior is vast, investigating methods to detection humans in single images, track humans across multiple frames and analyze human behaviors. We would like to mention at the outset of this work, that we do not tackle the problem of human detection or tracking. A number of methods exist in literature and the interested reader is directed to the reference[2] and the surveys[3)4)6)]. For our purposes, we track markers attached to the joints of the driver. The focus of this paper is the analysis of human operations or behaviors. Unlike most existing work, we do not assume the complete action is available at the onset of processing. Instead, we are interested in accurate prediction as early as possible in the evolution of an operation.

(a)



(b)

Fig. 1 Photos of the driving simulator and driver indicating six anatomical landmarks. The goal of this paper was to predict as early as possible what a driver intends to operate, before the operations are complete. (a) The locations of the six markers on the anatomical landmarks were used as input signals to the system. (b) The nine operation targets were; (1) navigation, (2) A/C, (3) left vent, (4) right vent, (5) gear box, (6) console box, (7) passenger seat, (8) glove box, and (9) rear-view mirror.

In general, the Hidden Markov Model (HMM) has been a popular technique for recognizing human behavior. It is useful for detecting patterns that indicate specified behaviors from a temporal sequence. Starner and Pentland[10] used HMMs to recognize the gestures in American Sign Language. Kahol et al.[5] employed the HMM based on human anatomy to recognize everyday human motion. In the area of driving behavior, Kuge et al.[7] and Oliver and Pentland[8] used a HMM to predict lane changing behavior as a primary driving task. However, it is difficult to apply HMMs for prediction. HMMs require the observation of a certain interval of the sequence to detect target behaviors. Such observation intervals delay predictions, which is detrimental to our goal of obtaining driver operations as quickly as possible.

In another prediction method, Salvucci[9] simulated a driving plan by modeling a knowledge-based cognitive architecture, ACT-R, to infer driver intent. The driver has the obvious purpose of controlling the vehicle in the environment and uses a consistent strategy to achieve this. Additionally, there are operations that result from a driver's needs. Therefore, it is possible to predict the operations by modeling mechanism of such needs based on the driver's profile, and the conditions inside and outside the car. However, it is merely an estimation of static probability of whether an operation is likely to take place and not an accurate prediction of when the driver will take the action. Consequently, it is important to effectively utilize lower-level information, including posture and movement. Additionally, Cheng et al.[1] present an approach for recognizing driver activities using a multi-perspective (i.e., four camera views) and multi-modal (i.e., thermal infrared and color) video-based system for robust and real-time tracking of important body parts.

3. OPERATION PREDICTION

At each time instant t, we wish to compute the probability

$p\ (o_m \mid d_t)$ that the current observed configuration of the driver's joint locations d_t originates from the operation o_m. The distribution of data in this feature space is complex. As a result, there are no simple classifiers that can used to learn the mapping between configurations and different operations (Fig. 2(a)), particularly early on in the execution of the action. By clustering the data space, the classification problem is reduced to a collection of simpler discrimination tasks, shown in Fig. 2(b).

The probability can then be computed using Bayes Theorem,

$$p\ (o_m \mid d_t) = \sum_j p\ (o_m, c_j \mid d_t) = \sum_j p\ (o_m \mid c_j, d_t)\ p\ (c_j \mid d_t) \quad \cdots\cdots\cdots\cdots (1)$$

where $\{c_j\}$ are the set of clusters estimated from the training data. We compute the probability $p\ (c_j \mid d_t)$ using per-cluster models, i.e. $p\ (c_j \mid d_t) = N\ (d_t \mid \mu_j, \Sigma_j)$ where $N\ (\cdot)$ is the multivariate Gaussian distribution function whose parameters (μ_j, Σ_j) are learnt from training data. We describe the model estimation process in the next section.

The probability $p\ (o_m \mid c_j, d_t)$ is, in turn, factored as,

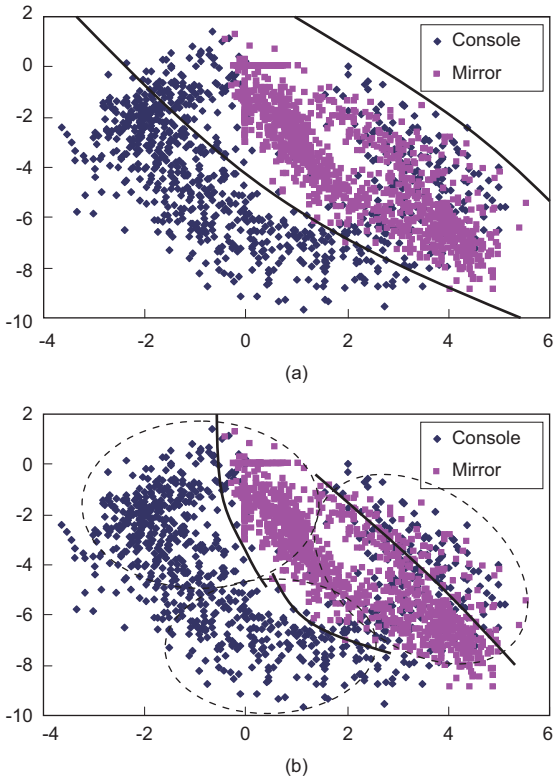


Fig. 2 (a) Two-class discriminant analysis.
(b) Discriminant analysis after cluster analysis.

$$p\ (o_m \mid c_j, d_t) = \frac{p\ (d_t \mid o_m, c_j)\ p\ (o_m \mid c_j)}{\Sigma_m p\ (d_t \mid o_m, c_j)\ p\ (o_m \mid c_j)} \quad \cdots\cdots\cdots (2)$$

The probability $p\ (o_m \mid c_j)$ is estimated by measuring the frequency of training data originating from o_m in c_j and $p\ (d_t \mid o_m, c_j) = N\ (d_t \mid \mu_{m,j}, \Sigma_{m,j})$. Once we evaluate $p\ (o_m \mid d_t)$ for all m, we predict that operation $o_m = r$ is under way if $p\ (o_m = r \mid d_t) \geq \tau > 0.5$, where τ is an empirical constant. The complete testing algorithm is summarized in Fig. 3.

4. MODEL BUILDING

Models for each operation are learnt from a corpus of training data. The configuration of the driver at each time instant is used to generate a feature vector, d_t. The design of this feature space is empirical and we justify its selection in Section 5. Each training sequence i is defined by a time-ordered collection of these vectors, $s_i = (d_0, \cdots, d_n)$. The set of sequences corresponding to each operation o_m defines the set S_{o_m}, $s_i \in S_{o_m}$, $\forall L\ (s_i) = o_m$, where $L\ (\cdot)$ provides the operation label for sequence s_i. The entire training set, containing many instances of the nine operations, is defined as $S = S_{o_1} \cup S_{o_2} \cup \cdots \cup S_{o_9}$. Instead of performing discriminant analysis directly on the data, we perform two intermediate steps. First, we divide S into k clusters, $c_1, \cdots, c_k$, using k-means clustering. k-means finds the centroid μ_j for each cluster c_j. Clustering in this way, before applying discriminant analysis on each cluster makes the classification task tractable. Second, since all operations commence from the same neutral onfiguration (hands on steering wheel), they are indistinguishable at the early stages of the operation. Thus, once clustering is performed, we weigh each data point d_t, by a weight determined by a monotonically increasing function $f\ (t)$, such as a linear, sigmoid or step function. Thus greater emphasis is place on more discriminant configurations that occur later on in each sequence.

Once clustering and data weighting is complete, multiple discriminant analysis is applied to each cluster to derive a mean $\mu_{m,j}$ and a covariance $\Sigma_{m,j}$ for operation o_m. The probability that a feature vector d_t^i is from operation o_m given its cluster membership c_j is computed as,

$$p\ (d_t^i \mid o_m, c_j) = \frac{1}{(2\pi)^{N/2} \left\| \Sigma_{m,j} \right\|^{1/2}}\ \mathrm{epx}\left(-\frac{1}{2} k\ (d_t^i \mid o_m, c_j)\right) \cdots (3)$$

where k is the Mahalanobis distance,

Training Algorithm

Given training data $\{d_t^i, o_{L(s_i)}\}$ where $d_t^i \in R^{16}$ is a feature vector from sequence i at time in instant t.

1. Clustering: Cluster data points $\{d_t^i\}$ using *k*-means algorithm. Learn a per-cluster model (μ_j, Σ_j).
2. Weighting: Weigh all data points $\{d_t^i\}$ using a temporal weighing function $f(t)$,e.g. linear, sigmoid or step function.
3. Parameter Estimation: Use multiple discriminant analysis to learn models $(\mu_{m,j}, \Sigma_{m,j})$ for each operation.

Testing Algorithm

Given an input configuration d_t.

1. Association: Compute the probability of association, $p(c_j|d_t)$ for all clusters $j = 1, \ldots, J$.
2. Discriminant Analysis: For each cluster j, compute $p(d_t|o_m, c_j)$, for $m = 1, \ldots, 9$ (for all nine operations).
3. Bayes Theorem: Use Bayes Theorem to compute $p(o_m|d_t) = \sum_j p(o_m, c_j|d_t)$
4. Decision: Select operation o_m if $p(o_m|d_t) \geq \tau > 0.5$

Fig. 3 Training and testing algorithms.

$$k(d_t^i \mid o_m, c_j) = (d_t^i - \mu_{m,j})^T \Sigma_{m,j}^{-1} (d_t^i - \mu_{m,j}) \cdots\cdots (4)$$

The training algorithm is summarized in **Fig. 3**.

5. FEATURE SPACE SELECTION

We selected a feature space that consisted of a combination of location and velocity vectors. We evaluated different feature spaces on data sets of five subjects, collected during driving in a driving simulator. In the first experiment, model creation and testing were executed for each driver using only instantaneous location information of each point in the configuration. The data set for each driver, which included 450 operations (90 operations each for 5 drivers), was divided into four blocks and leave-one-block-out cross-validation was performed. **Figure 4** shows the recognition rate in the best case, worst case and on average within five subjects over normalized time from the start of the operation. At the half-point of operation (50% operation), the recognition rate is 85%, and even in the worst case, it reaches over 75%, indicating that potential for predicting driver operation.

We then evaluated the performance on the various combinations of the movement vectors for six joints. Adding the movement vectors of the left elbow and the left wrist to six joint positions, the best performance was obtained, shown in **Fig. 5**(a). The recognition error rates are shown in **Fig. 5**(b). At 50% operation, the average recognition rate increases to 94%, improving 9% from the rate in the case of six joints. On the other hand, the maximum recognition error rate slightly increases for the first period of the operations by adding two movement vectors, shown in **Fig. 6**. Thus, the feature space used in this paper is

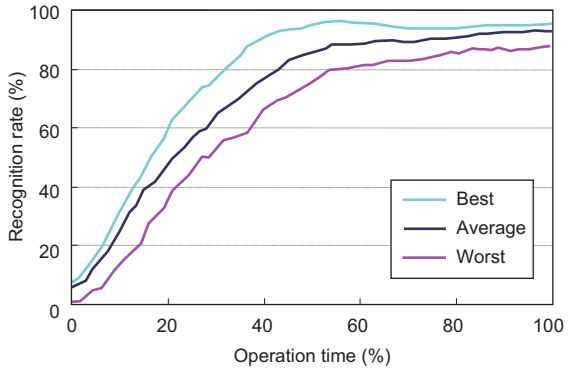
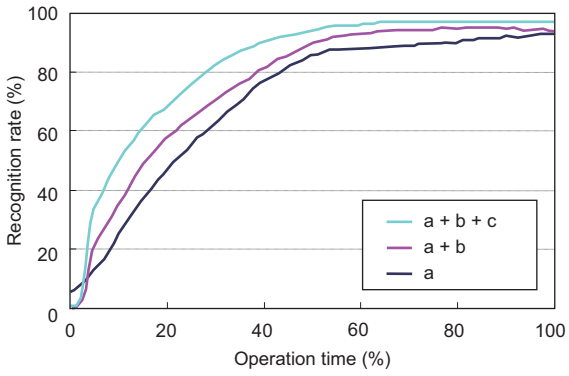
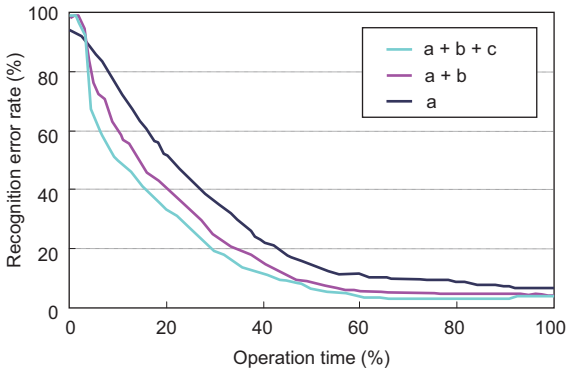


Fig. 4 Result of estimation directly over configuration space for a dataset of 450 operations.

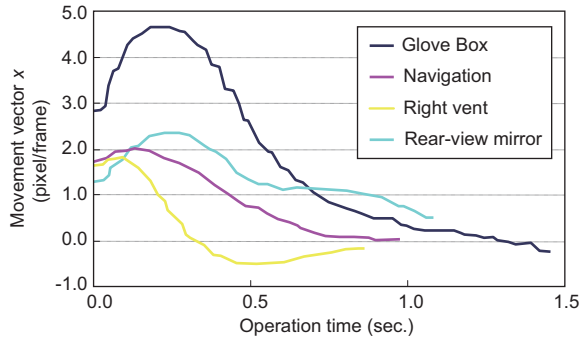


(a) Recognition rate

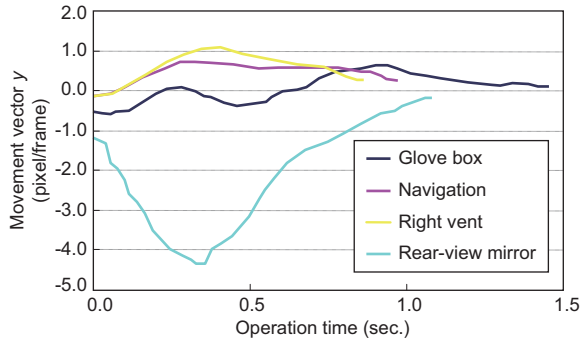


(b) Recognition error rate

Fig. 5 The influence of movement vectors on the recognition and error rate. The curves show the influence of 'a' (configuration vector), 'b' (movement vector for the left wrist), and 'c' (movement vector for the left elbow).



(a) Movement vector *x*



(b) Movement vector *y*

Fig. 6 Average movement vector of the left wrist for four operations. Movement is a strong cue in discriminating between different operations shown by the trajectories on the *x*-axis (a) and *y*-axis (b).

$$d_t = [x_0, y_0, \cdots, x_5, y_5,, u_4, v_4, u_5, v_5]^T \in R^{16},$$

where x_j and y_j are the image locations, u_j and v_j are image velocities.

6. RESULTS

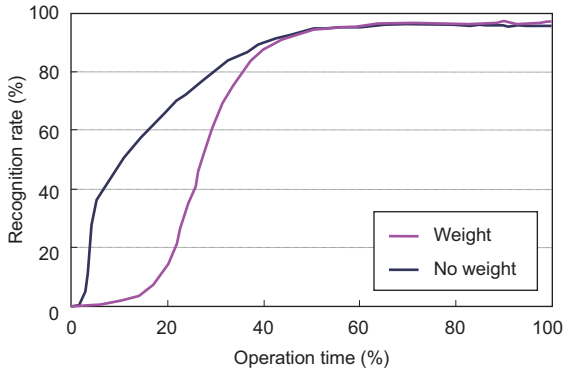
The system shows significant accuracy in predicting operations as is shown in **Fig. 9**. The top row shows instances at which actions are correctly predicted for eight operations. Based on the location and velocity information, the proposed method was applied to predict nine frequently executed operations inside a vehicle, running at an average rate of 172 frames per second. For five subjects, the method achieves an average prediction accuracy of 90% with a false positive rate of 1.4% at half the operation duration. The bottom row shows the eventual destination configurations for the three operations.

Figure 8 shows the likelihood of each operation at each time instance in a sequence where the driver reaches for the navigation console. At 25% of the operation, there exists confusion between similar actions like adjusting the A/C. As the sequence progresses however, the algorithm is able to predict the operation accurately before half the operation duration. **Table 1** shows the individual prediction rates for all nine operations at half the duration of the operation. The highest rates were observed for the operations which were most spatially separated, i.e. adjusting the rear-view mirror, opening the center console box, and opening the glove compartment. The lowest accuracy was recorded for adjusting the air conditioner, since there are many similar actions in the set of nine operations.

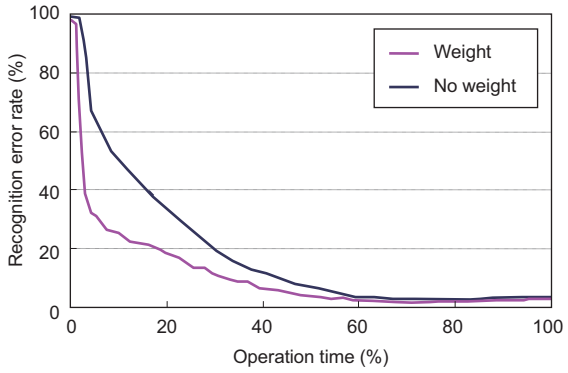
Figure 10(b) shows the comparison with the result by an HMM-based approach. To create the HMM for prediction, every operation sequence was divided into four sequences and the HMM was trained on these segmented sequences (the number of state and the number of possible observations in sequence were decided by experiment). The proposed method has an inherent advantage over the HMM for predicting driver operations. The target operations are simple and short actions not like gestures for communication so that it may be difficult to represent the differences among equipment as sequence of state transition on the HMM.

Table 1 Individual prediction results.

	A/C	Console	Glove	Navi	L Vent	R Vent	P Seat	Mirror	Gear	Steer
A/C	82.5	0.0	0.0	4.5	0.7	8.0	1.0	0.0	0.0	3.1
Console	1.2	94.8	0.0	0.0	0.0	0.0	0.0	0.4	2.8	0.8
Glove	1.2	0.0	95.6	0.0	1.6	0.0	1.2	0.0	0.0	0.4
Navi	5.0	0.0	0.0	92.9	1.1	0.4	0.4	0.0	0.0	0.4
L Vent	1.1	0.0	0.0	0.8	97.7	0.0	0.4	0.0	0.0	0.0
R Vent	1.8	0.0	0.0	0.4	0.4	95.8	0.0	0.0	0.0	1.8
P Seat	0.0	0.8	0.0	0.0	0.0	0.0	96.9	98.0	1.9	0.4
Mirror	1.2	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.4
Gear	0.0	0.7	0.0	0.0	0.0	0.0	5.0	0.0	93.9	0.4



(a) Recognition rate



(a) Recognition error rate

Fig. 7 Influence of weighting on the recognition and error rates.

We also evaluated the benefit of using the weight for training, where the discrete variable of the step is set to 0.3. The error rate decreases for the first period and the maximum error rate also decreases to 11%. **Figure 7**(a) shows the recognition rates. When this weighting scheme is used, increasing is delayed, but at 50% operation the recognition rate reaches the same percentage as in the case of no weight.

Figure 10(a) graphs the true-positive rate versus false-positive (or false alarm) rate at 50% operation. The result with the weight is significantly better than that of no weight. As the true-positive rate is 90%, the false-positive rate is 1.4% for weight and 8.4% for no weight. It describes that our method can decrease the false-positive rate without making the true-positive rate worse.

7. DISCUSSION

The proposed method can quickly and accurately predict nine operations. The method achieves 90% true-positives with 1.4% false-positives at half of the operation duration, running at an average rate of 172 frames per second. Using labeled training data, we use clustering following applying by multiple discriminant analysis on each cluster to model configurations of each object. During testing, we predict which operation is most likely to be under execution given the instantaneous configuration by evaluating membership in each operation set.

Future Work: We are currently developing a marker-less detection and tracking method. This tracking method will be incorporated into the prediction method to evaluate the complete prediction system, and we will evaluate how the biases of the tracking method affect prediction performance. Additionally, we have assumed that drivers grasp the steering wheel at the start of operations. In practice, however, operations may commence at any posture. We are now extending the method to manage natural initiations of operations. Finally, we will investigate customized training methods, where we create an individual model for each

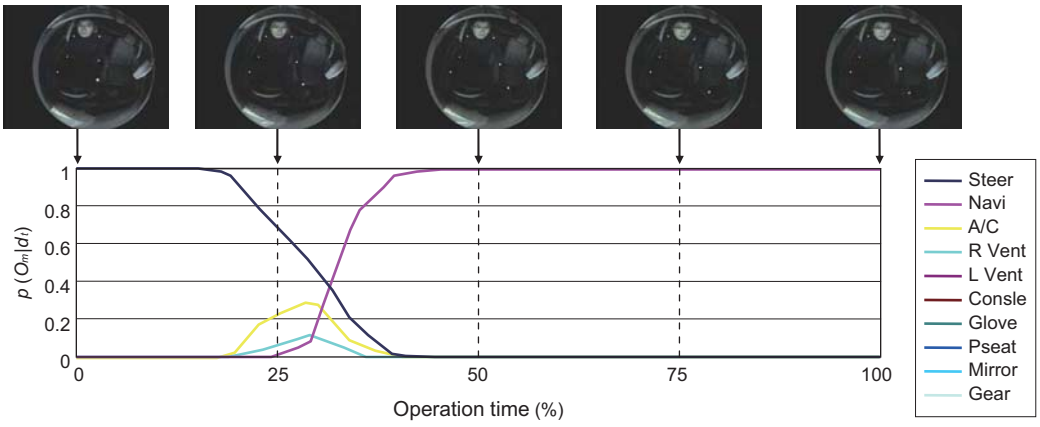
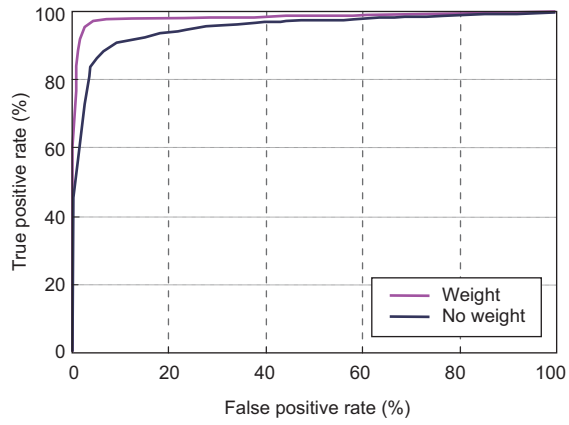


Fig. 8 Predicting “accessing the navigation console” at different instances. There was some initial confusion between operations that was quickly resolved over time.

Fig. 9 Images at the time of prediction for eight driver operations. Top: Images at the time when the proposed approach outputs the correct answer. Bottom: final configuration of the operation.



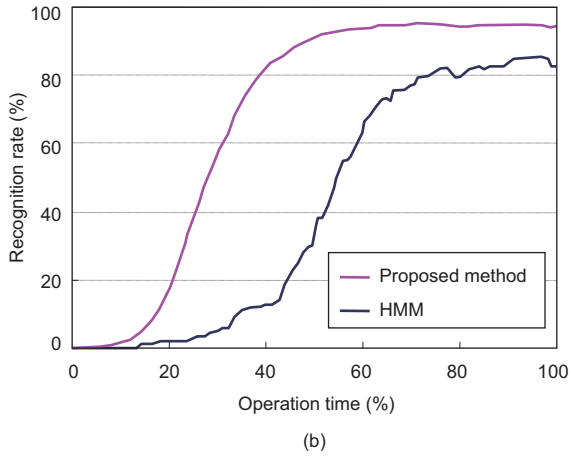


Fig. 10 (a) Comparison between True Positives and False Positives. (b) Comparison using an HMM-based approach.

driver. This strategy will be tested in real vehicles. The system will extract training sequences by sensing the state of equipment's interface (e.g. switch, dial, and touch sensor), showing the start time and end time of operations. We can utilize a prior model that the system then customizes by gathering data.

REFERENCES

1) S. Y. Cheng, S. Park, and M. M. Trevedi. Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis, 2007. CVIU.
2) A. Datta, Y. Sheikh, and T. Kanade. Linear motion estimation for systems of articulated planes. IEEE International Coference on Computer Vision and Pattern Recognition, 2008.
3) D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramaman. Computational studies of human motion: Part 1, tracking and motion synthesis. Foundations and Trends in Computer Graphics and Vision, 2006.
4) D. Gavrila. The visual analysis of human movement: A survey. Computer Vision and Image Understanding, 1999.
5) K. Kahol, P. Tripathi, and S. Panchnathan. Recognizing every human movements through human anatomy based coupled hidden markov models. IEEE International Conference on Automatic Face and Gesture Recognition, 2004.
6) T. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, 2006.
7) O. S. N. Kuge, T. Yamamura and A. Lie. A driver behavior recognition method based on a driver model framework. SAE Trans., 109:469-476, 2000.
8) N. Oliver and A. Pentland. Driver behavior recognition and prediction in a smartcar. Proceeding of SPIE Aerosense-Enhanced and Synthetic Vision, page 280-290, 2000.
9) D. Salvucci. Inferring driver intent: A case study in lane-change detection. Proceeding of the Human Factors Ergonomics Society, 2004.
10) T. Starner and A. pentland. Visual recognition of sign language using hidden markov model. International Workshop on Automatic Face and Gesture recognition, 1995.

＜著　者＞

伊藤　隆文
（いとう　たかふみ）
基礎研究所
ドライバの状態検出技術の研究に従事

金出　武雄
（かなで　たけお）
カーネギーメロン大学
ウイタカー記念全学教授
主な経歴として 1992 年-2001 年カーネギーメロン大学ロボティックス研究所所長，日本において 2001 年，産業技術総合研究所でデジタルヒューマンラボ（現デジタルヒューマン工学研究センター）を興し，所長を兼務，2010 年より特別フェロー．
コンピュータビジョン，ロボット，自律走行車，マルチメディア等の研究に従事．