

Recognition of driving postures by combined features and random subspace ensemble of multilayer perceptron classifiers

Chihang H. Zhao · Bailing L. Zhang ·
Xiaozheng Z. Zhang · Sanqiang Q. Zhao ·
Hanxi X. Li

Received: 13 June 2011 / Accepted: 22 June 2012 / Published online: 11 July 2012
© Springer-Verlag London Limited 2012

Abstract Human-centric driver assistance systems with integrated sensing, processing and networking aim to find solutions for traffic accidents and other relevant issues. The key technology for developing such a system is the capability of automatically understanding and characterizing driver behaviors. This paper proposes a novel driving posture recognition approach, which consists of an efficient combined feature extraction and a random subspace ensemble of multilayer perceptron classifiers. A Southeast University Driving Posture Database (SEU-DP Database) has been created for training and testing the proposed approach. The data set contains driver images of (1) grasping the steering wheel, (2) operating the shift lever, (3) eating a cake and (4) talking on a cellular phone. Combining spatial scale features and histogram-based features, holdout and cross-validation experiments on driving posture classification are conducted, comparatively. The experimental results indicate that the proposed combined feature extraction approach with random subspace ensemble of multilayer perceptron classifiers outperforms the two individual feature extraction approaches. The experiments also suggest that talking on a cellular phone is the most difficult posture in classification among the four

predefined postures. Using the proposed approach, the classification accuracy on talking on a cellular phone is over 89 % in both holdout and cross-validation experiments. These results show the effectiveness of the proposed combined feature extraction approach and random subspace ensemble of multilayer perceptron classifiers in automatically understanding and characterizing driver behaviors toward human-centric driver assistance systems.

Keywords Driver assistance system · Random subspace ensemble · Driving postures · Multilayer perceptron classifier · Combined features

1 Introduction

Driving is stressful, which requires intensive cognitive processing on the part of the driver, as well as carefully performing maneuvers. To alleviate the driving stresses, recent researches on vehicle automation started to focus on developing human-centric driver assistance systems. Usually, such a system includes integrated sensing, processing and networking and aims to find solutions to traffic problems (e.g., traffic accidents, traffic congestion, etc.). Automatic understanding and characterizing driver behaviors is one of the key elements in human-centered driver assistance systems. A driver's behavior generally indicates his or her driving capability, such as attention, fatigue levels, or other unpredictable and distractive factors. Unsafe behaviors, such as fatigue, eating and talking on a cellular phone while driving, reduce the driver's alertness and make the driver more prone to road accidents. Nadeau et al. [1] carried out an epidemiological study on two large cohorts, namely users and non-users of cellular phones. It was found that the adjusted relative risks for heavy users of

C. H. Zhao (✉)
College of Transportation, Southeast University,
Nanjing 210096, People's Republic of China
e-mail: chihangzhao@seu.edu.cn

B. L. Zhang
Department of Computer Science and Software Engineering,
Xi'an Jiaotong-Liverpool University, Suzhou 215123,
People's Republic of China

X. Z. Zhang · S. Q. Zhao · H. X. Li
Queensland Research Laboratory, National ICT Australia,
Brisbane, QLD 4111, Australia

cellular phones are at least twice of clean drivers. Hence, correctly recognizing and alerting unsafe driving behaviors can significantly improve road safety. The development of an effective driver behavior recognition system is twofold, (a) driver's postures or maneuvers need to be efficiently described, and (b) those posture descriptors should be reliably classified.

Vision sensors can capture physical information about a driver's postures in an untethered manner. In the last decade, many new techniques have been proposed for the recognition of some simplified driving postures, such as moving forward, turning left and turning right. Oliver and Pentland [2] proposed a machine learning framework for modeling and recognizing a driver's movements. Using graphical models (Hidden Markov Models and Coupled Hidden Markov Models), their system emphasized on the effect of contexts on driver's performance. Liu et al. [3] developed a vision system that tracks driver's faces and estimates face poses while driving using yaw orientation angles. Eren et al. [4] presented a stereo vision system for predicting driver's face poses using principal components analysis (PCA). Sekizawa et al. [5] designed an intelligent system of modeling driver's collision avoidance behaviors at the moment that the vehicle ahead is brought to a sudden halt. Watta et al. [6] presented a vision system that recognizes the directions the driver is looking using a nearest neighbor-neural network classifier. To address the problem of poor and unstable illumination in driving images captured by a color camera, Kato et al. [7] developed a far infrared camera system. The feature points of nose, mouth and ears were used to estimate the directions of the driver's gaze. Cheng et al. [8] introduced a video-based system for recognizing the driver's body orientation. Multiple cues were used including thermal and visual color images, as well as steering wheel angle data collected by Controller Area Network (CAN) bus of the vehicle. Demirdjian and Varri [9] used an infrared time-of-flight (TOF) camera to estimate the location and orientation of a driver's head, torso and limbs (i.e., arms and hands). Cheng and Trivedi [10] suggested a commercial motion-capture system to identify a driver's body pose that places retro-reflective markers on the driver's bodies. Yang et al. [11] presented an ECG system to monitor a driver's sitting postures, that is, forward movement, back movement, right back movement and left back movement. In [12], a modified Histogram of Oriented Gradients (HOG) feature descriptor and a support vector machine were used to classify occupants of the front seats as driver, passenger or no one.

Most of the above research works about driver activities recognition focused on the detection of the driver's face direction, head orientation and gaze, etc. Recently, many researches have been done on human pose and gesture recognition. For example, hidden Markov model (HMM)

was proposed for hand localization, hand tracking and gesture spotting as the preprocessing step for hand gesture recognition in [13, 14]. In [15], hand candidate regions are located on the basis of skin color and motion. The centroids of the moving hand regions are connected to produce a hand trajectory, which are divided into real and meaningless segments, and the combined features vectors are extracted from weighted location, angle and velocity. The framework for segmenting and tracking skin objects from signing videos consists of a skin color model and a skin-object tracking module. The skin color model was built on the combination of Support Vector Machine learning and region segmentation, and the tracking module predicts occlusions among of skin objects using Kalman filter. In [16], dynamic Bayesian network (DBN) was proposed to recognize hand gestures in a continuous video stream, which was also preceded by steps of skin color extraction and modeling, and motion tracking. An other approach of using a skin color for hand gesture recognition was presented in [17], which was based on Gabor filters and support vector machine (SVM). Nevertheless, fewer efforts have been made toward the recognition of driver postures comparing with other kind of human postures recognition. Only Harini et al. [18, 19] proposed an agglomerative clustering and a Bayesian eigen-image classifier to recognize two types of a driver postures (safe and unsafe) using a side-mounted camera to capture the driver's profile. How to characterize more realistic driving postures, however, is still a challenging problem.

A crucial step in developing image-based driver posture recognition is to extract suitable features from the driver's images and characterize differences between different driving postures. In this paper, we proposed an efficient combined feature extraction approach, which is constructed based on spatial scale features description and histogram-based features description, and a random subspace ensemble of multilayer perceptron classifiers was then exploited in the classification of the driving postures. The rest of the paper is organized as follows. In Sect. 2, the background of SEU driving postures data acquisition and normalization is outlined. In Sect. 3, spatial scale feature extraction, histogram-based feature extraction and combined feature extraction are introduced. The random subspace ensemble of multilayer perceptron classifiers is presented in Sect. 4. Section 5 details the experiments and discusses the classification results for the driving postures. Section 6 draws the conclusions.

2 Driving postures data acquisition and normalization

To test the proposed driving posture recognition approach, a driving posture data set (Southeast University Driving Posture database, or hereafter SEU-DP database for short)

has been created. We used a side-mounted Logitech C905 CCD camera to capture driving posture images. Images of 20 individuals (10 male and 10 female) were acquired in four different driving postures, that is, grasping the steering wheel, changing gear, eating a cake and talking on a cellular phone. The images were produced under day lighting conditions, when the car was in an outdoor parking lot. The SEU-DP data set consists of four driving postures. The images are all of resolution 640×480 pixels. Figure 1 shows several samples of SEU-DP data set with the four different driving postures.

In order to address the problem of illumination variations in images of SEU-DP data set, we adopted a normalization technique, homomorphic filter (HOMOF) [20], to enhance the image quality. The objects of interest in the driving images are the skin-like regions, mainly from the driver's head and two hands. In fact, human skin tones have similar chromatic properties regardless of race. Hence, skin color detection can be fairly robust under natural illumination conditions. The classification of color pixels into skin tones and non-skin tones is performed by working in the normalized RGB space. An RGB triplet (r, g, b) with values for each primary color between 0 and 255 is normalized into the triplet (r', g', b') by using

$$r' = \frac{255r}{r+g+b}, g' = \frac{255g}{r+g+b}, b' = \frac{255b}{r+g+b}. \quad (1)$$

The normalized color (r', g', b') is classified as a skin color if it lies within the skin region in the normalized RGB space. The thresholding rules are [19]

$$\begin{cases} r' > 95, & g' > 45, & b' > 20 \\ \max\{r', g', b'\} - \min\{r', g', b'\} > 15. \\ r' - g' > 15, & r' > b' \end{cases} \quad (2)$$

Figure 2 shows the skin color segmentation result of Fig. 1d without HOMOF. It includes some undesired background of the vehicle's interior. Figure 3 shows the skin color segmentation results of Fig. 1a–d preprocessed by HOMOF, none of which has unwanted background. Therefore, HOMOF preprocessing step can increase the robustness of the skin color segmentation.

3 Feature extraction of driving postures

Human bodies appear in a wide range of configurations and tend to be self-occluded. Recognizing human body poses then becomes a challenging vision task. The illumination robustness requirements of vehicle devices further complicate this challenge in body posture recognition. In this section, we show the promise of two different feature expression methods to build the driver's posture feature vectors. The first is the extraction of histogram-like

Fig. 1 Example images of SEU-DP data set. **a** Grasping the steering wheel, **b** operating the shift lever, **c** eating a cake, **d** talking on a cellular phone





Fig. 2 Skin color segmentation of Fig. 1d without HOMOF

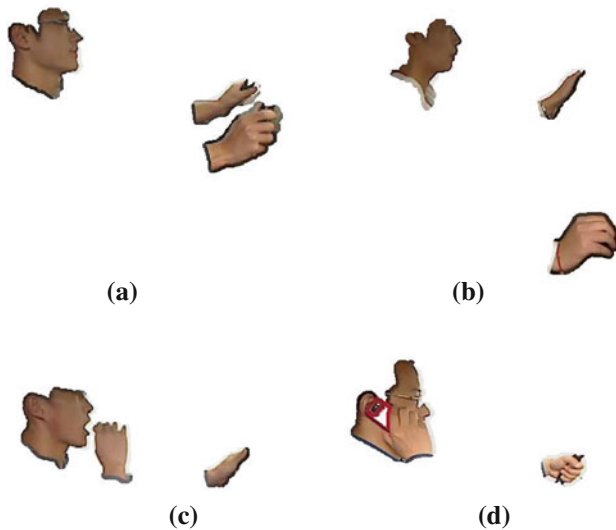


Fig. 3 Skin color segmentation of images in Fig. 1 preprocessed using HOMOF. **a** Grasping the steering wheel, **b** operating the shift lever, **c** eating a cake, **d** talking on a cellular phone

features that emphasize on edge information, while second is using the exposed skin spatial position of the driver's head, right hand and left hand. An appropriate combination of them can make the feature expression more efficient.

3.1 Histogram-based feature extraction by Pyramid Histogram of Oriented Gradients (PHOG)

Histogram of Oriented Gradients (HOG) [21] is a widely used local edge descriptor in computer vision applications. It has been successfully applied to a number of object detections such as human detection. When HOG encodes the gradient orientation of one image patch, however, it does not take into account the location of this orientation in the patch. Consequently, it is not descriptive to the spatial property of the underlying structure of the image patch. To alleviate this issue, Pyramid Histogram of Oriented Gradients (PHOG) [22] was proposed to encode the spatial property of the local shape while representing an image by HOG. The spatial information is represented by tiling the

image into regions at multiple resolutions, based on spatial pyramid matching. The image is divided into a sequence of increasingly finer spatial grids, which repeatedly double the number of divisions along each axis. The number of points in a grid cell at a certain level is then recorded as the sum of point numbers over the corresponding four grid cells in the next finer level. In this way, the representation forms a pyramid HOG. A single feature vector is formed by concatenating all of the HOG in different level and grid cells. A weighted sum over the histogram intersections of PHOG is used to match the PHOGs of two images.

In the implementation, we set the number of levels to $L = 3$ to prevent over fitting [22]. The PHOG is normalized to sum to unity to balance the images with denser and sparser edges. The PHOG descriptors are computed from the gray-level images of skin color segmented images as shown in Fig. 3. Taking Fig. 3c as an example, the PHOG processing is illustrated in Fig. 4. Figure 4a shows the gray-level image of Figs. 3c and 4b–d show HOGs of the image at Level 1 (entire image), Level 2 (intermediate 2×2 image regions) and Level 3 (the finest 4×4 image regions), respectively. It is demonstrated that the location property of the segmented image can be encoded in the PHOG with different grid levels.

3.2 Spatial scale feature extraction approach

The Canny edge detection algorithm [23] is applied to the skin color segmentation results to implement edge detection in images in SEU-DP data set. Let $E(x, y)$ be the binary image from the edge detection. The isolated small regions are deleted to form a refined binary image $M(x, y)$. As shown in Fig. 5, $M(x, y)$ consists of three major connected skin regions, that is, the driver's head, the left hand and the right hand.

For the i th connected region, let $x_{\min}^i, x_{\max}^i, y_{\min}^i$ and $y_{\max}^i$ ($1 \leq i \leq 3$) be the minimum and maximum values of the x -coordinates, the minimum and maximum values of the y -coordinates, respectively. The approximate centric coordinates of the i th connected region in the resulting edge image are calculated as

$$\begin{cases} x_i = \frac{x_{\max}^i - x_{\min}^i}{2} \\ y_i = \frac{y_{\max}^i - y_{\min}^i}{2} \end{cases} \quad (3)$$

Therefore, (x_1, y_1) , (x_2, y_2) and (x_3, y_3) denote the approximate centric positions of the driver's head region, the driver's left-hand region and the driver's right-hand region in the resultant edge image. The approximate centric distance l_1 between the driver's left-hand region and the driver's head region in the resultant edge image is calculated as

Fig. 4 Schematic illustration of PHOG. **a** Example image, **b** $L = 1$, **c** $L = 2$, **d** $L = 3$

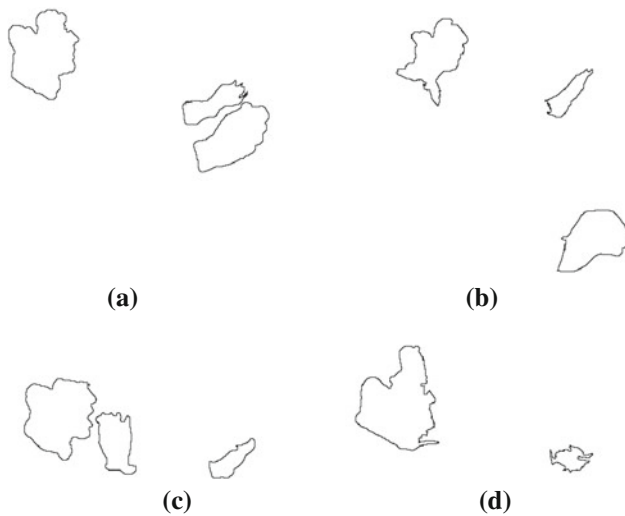
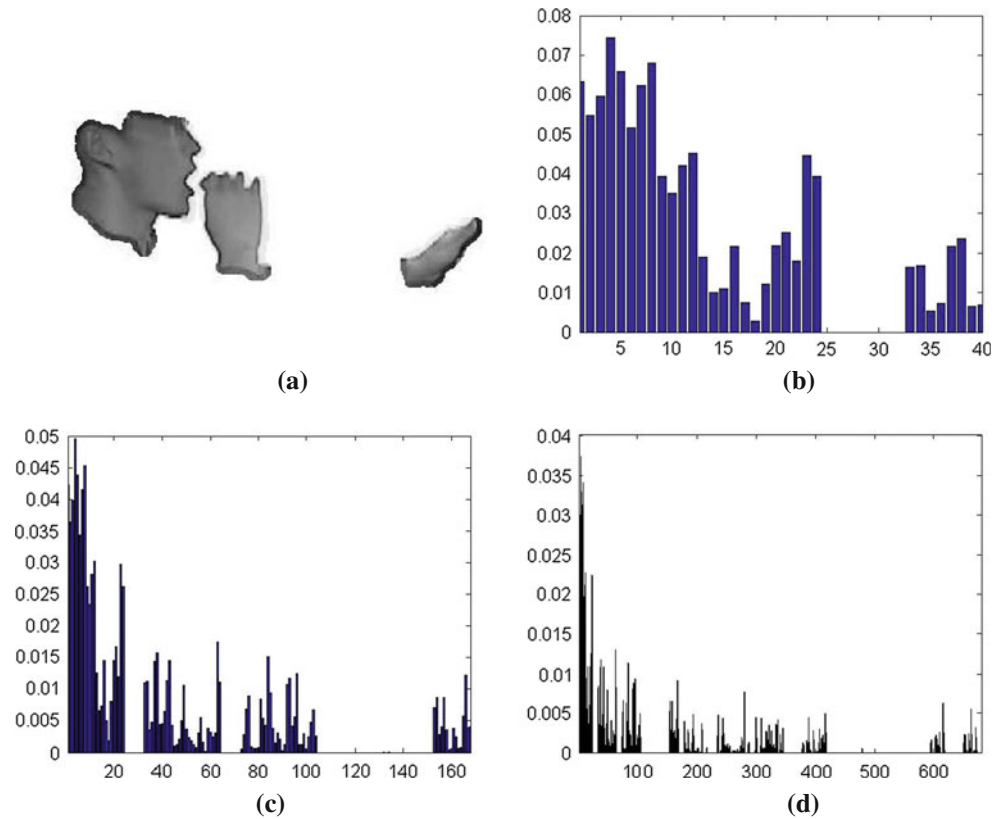


Fig. 5 Results of edge detection and refinement of the four driving postures in Fig. 3. **a** Grasping the steering wheel, **b** operating the shift lever, **c** eating a cake, **d** talking on a cellular phone

$$l_1 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (4)$$

The approximate centric distance l_2 between the driver's left-hand region and the driver's right-hand region in the resultant edge images is calculated as

$$l_2 = \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2} \quad (5)$$

The approximate centric distance l_3 between the driver's head region and the driver's right-hand region in the resultant edge images is calculated as

$$l_3 = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} \quad (6)$$

It is observed that different driving postures correspond to different centric distances l_1 , which can be used to differentiate driving postures. To make these descriptions robust to scale changes, two ratios of the three distances are calculated and used. The spatial scale ratio of the approximate centric distances l_2 and l_1 is denoted as v_1 , and the spatial scale ratio of the approximate centric distances l_3 and l_1 is denoted as v_2 . The two descriptors (v_1 and v_2) of the images in SEU-DP data set are calculate as

$$\begin{cases} v_1 = l_2/l_1 \\ v_2 = l_3/l_1 \end{cases} \quad (7)$$

Each feature vector extracted from above two different methods characterizes certain aspects of the driving postures image. The joint exploitation of histogram-based feature extraction and spatial scale-based feature extraction is often necessary to provide a more comprehensive description for a driving posture classification system with higher accuracy. The difficulties of two type feature aggregation lie in high dimensionalities of the image

feature vectors. However, with random subspace ensemble of multilayer perceptron classifiers, which will be elaborated in the following section, this problem can be implicitly resolved due to its dimension reduction capability.

4 Random subspace ensemble of multilayer perceptron classifiers

A classifiers ensemble can individually train a set of classifiers and appropriately combine their component decisions [24, 25]. Generally, classifier ensembles are able to improve classification performances. There are many ways to form a classifier ensemble. A mainstream methodology is to train the homomorphous ensemble members using different subsets of the training data. It is usually implemented by resampling (bagging) [26] and reweighing (boosting) [27] the available training data. Different classifier solutions can be applied in ensemble learning. In this paper, we choose the neural classifiers as the base learners based on the simple consideration that a simple three-layer back propagation neural network (BPNN) can approximate any continuous function, given sufficient numbers of middle-layer units and the evidence that a BPNN can perform superbly in classification [31].

The multilayer perceptron (MLP) trained with the back-propagation algorithm has been successfully applied to many classification problems in the field of bioinformatics [28–30]. In a typical MLP, a set of source nodes forms the input layer, one or more hidden layers of computation nodes, and a layer of output nodes. Input–output mappings are constructed and their characteristics are determined by the weights assigned to the connections between the nodes in two adjacent layers. Changes in weights change the input-to-output behaviors of the network. To train an MLP, a derivable criterion (e.g., mean squared error) is usually optimized using a gradient-descent-based back-propagation algorithm [31].

When using multiple MLP classifiers, the inconsistency from different MLPs outcomes can be properly exploited to increase the classification performance via an ensemble strategy. To achieve the performance enhancement, a critical requirement in constructing MLP ensembles is to diversify the components of the ensemble. In this paper, we focus on MLP ensembles based on the random subspace (RS), a successful ensemble generation technique [32, 33]. The major procedures of RS are as follows. For a p -dimensional training set, an p^* -dimension feature vector ($p^* < p$) is randomly selected with uniform distribution. Thus, the data of the original d -dimensional training set are transformed to the selected p^* -dimensional subspace.

The resulting subset of feature vectors is then used to train a suitable base classifier. This process repeats R times, such that R base classifiers are trained on different subsets of feature vectors. The resulting R classifiers are combined by using weighted majority voting (i.e., Adaboosting). The details are as follows.

Considering a training set $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$, each training sample $\mathbf{X}_i$ is a p -dimensional feature vector, $\mathbf{X}_i = \{x_{i1}, x_{i2}, \dots, x_{ip}\}$, ($i = 1, \dots, n$). p^* ($< p$) feature elements are randomly selected from the original p -dimensional feature vector $\mathbf{X}_i$ to construct a new p^* -dimensional feature vector. The original training sample set $\mathbf{X}$ becomes $\mathbf{X}^r = \{\mathbf{X}_1^r, \mathbf{X}_2^r, \dots, \mathbf{X}_n^r\}$. Each training sample in $\mathbf{X}^r$ consists of p^* feature elements, that is, $\mathbf{X}^r = \{\mathbf{X}_1^r, \mathbf{X}_2^r, \dots, \mathbf{X}_{p^*}^r\}$, ($i = 1, \dots, n$). Feature element x_{ij}^r ($j = 1, \dots, p^*$) is randomly selected with uniform distribution. Then, R classifiers are constructed in the random subspace $\mathbf{X}^r$. Finally, these classifiers are aggregated using the weighted majority voting rule (i.e., Adaboosting). This RS procedure is formally expressed as follows:

Step 1: Repeat for $r = 1, 2, \dots, R$.

- (1) Select a p^* -dimension random subspace $\mathbf{X}^r$ from the original p -dimension feature space $\mathbf{X}$. Denote each p^* -dimensional feature vector as x^r .
- (2) Construct a classifier C^r (x^r) (with a decision boundary $C^r(x^r) = 0$ in $\mathbf{X}^r$).
- (3) Compute the combined weights

$$c_r = \frac{1}{2} \log \left(\frac{1 - \text{err}_r}{\text{err}_r} \right), \quad (8)$$

where $\text{err}_r = \frac{1}{n} \sum_{i=1}^n \omega_i^r \xi_i^r$. $\xi_i^r = 0$, if $\mathbf{X}_i$ is classified correctly, and $\xi_i^r = 1$, otherwise. Let y_i , ($i = 1, \dots, N$) denote the class label, and the function $f(X_i)$ with values ± 1 is in a given class. Weights are initialized as $\omega_i^1 = 1/N$, ($i = 1, \dots, N$) and updated using

$$\omega_i^{r+1} = \frac{\exp(-F_r(X_i) \cdot y_i)}{Z_{r+1}} \quad (9)$$

where $F_r(X_i) = \sum_{s=1}^R c_s f_s(X_i)$, $Z_{r+1} = \sum_{i=1}^N \exp(-F_r(X_i) \cdot y_i)$.

Step 2: Combine classifiers $C^r(x)$, $r = 1, 2, \dots, R$, by the weighted majority vote with weights c_r to a final decision rule

$$\beta(x) = \arg \max_{y \in \{-1, 1\}} \sum_r \delta_{\text{sgn}(C^r(x)), y}, \quad (10)$$

where $\delta_{i,j} = 1$, if $i = j$, $\delta_{i,j} = 0$ otherwise. $y \in \{-1, 1\}$ is a decision (class label) of the classifier.

5 Experiments

Two standard experimental procedures, holdout strategy and cross-validation, were used to compare the performances of the histogram-based feature extraction by PHOG, the spatial scale-based feature extraction and the combined feature extraction. In our experiments, the MLP classifier was configured as a structure with one hidden layer. The activation functions for hidden and output nodes are logistic sigmoid function and linear function, respectively. We experimented with MLP classifiers with 20 units in the hidden layer and 4 linear units representing the class labels. The MLP network was trained using conjugate gradient learning algorithm for 500 epochs. In the holdout experiment, certain amounts of driving posture feature vectors extracted from SEU-DP data set (shown in Fig. 6) were reserved for testing and the rest were used for training. In k -fold cross-validation, the SEU-DP data set is partitioned into k subsets. Among them, one subset was retained for testing the classification model and the remaining $k - 1$ subsets were used for training the classification model. The cross-validation experiments were then repeated k times, with all of the k subsets used exactly once as the validation data set. The k experiment results from the folds were then averaged to produce a single classification rate to report.

5.1 Holdout experiments

Holdout experiments were based on randomly splitting of feature vectors of driving postures into a training subset (80 % of feature vectors of driving postures extracted from

images in SEU-DP data set) and a test subset (20 % of feature vectors of driving postures extracted from images in SEU-DP data set). With the holdout experiment strategy, only the test subset was used to estimate the generalization error. We repeated the holdout experiment 100 times by randomly splitting dividing vectors of driving postures and recorded the classification results. For each random testing, the same set of training and testing are applied to the above three feature extraction approaches and their classification performances are simultaneously compared.

The results of classification rate for driving postures are displayed as bar plots in Fig. 7a and as box plots in Fig. 7b. They are the averaged classification results from 100 random splits of feature vectors extracted from SEU-DP data set. The average classification accuracies of histogram-based feature extraction by PHOG, spatial scale feature extraction and combined feature extraction are 93.5, 88.13 and 94.75 %, respectively. Figure 7 shows that combined feature extraction with random subspace ensemble of MLP classifiers offers the best performance in the holdout experiments among the tested three feature extraction methods.

Confusion matrix was used to further measure the classification performance regarding the information about actual and predicted classifications acquired. The confusion matrix is a square matrix that represents the proportion of driving postures from one class classified into each of the four possible classes. In the holdout experiment, the confusion matrix that summarizes the detailed performance of combined feature extraction with random subspace ensemble of MLP classifiers is shown in Table 1. In the confusion matrix, the rows and columns indicate true and

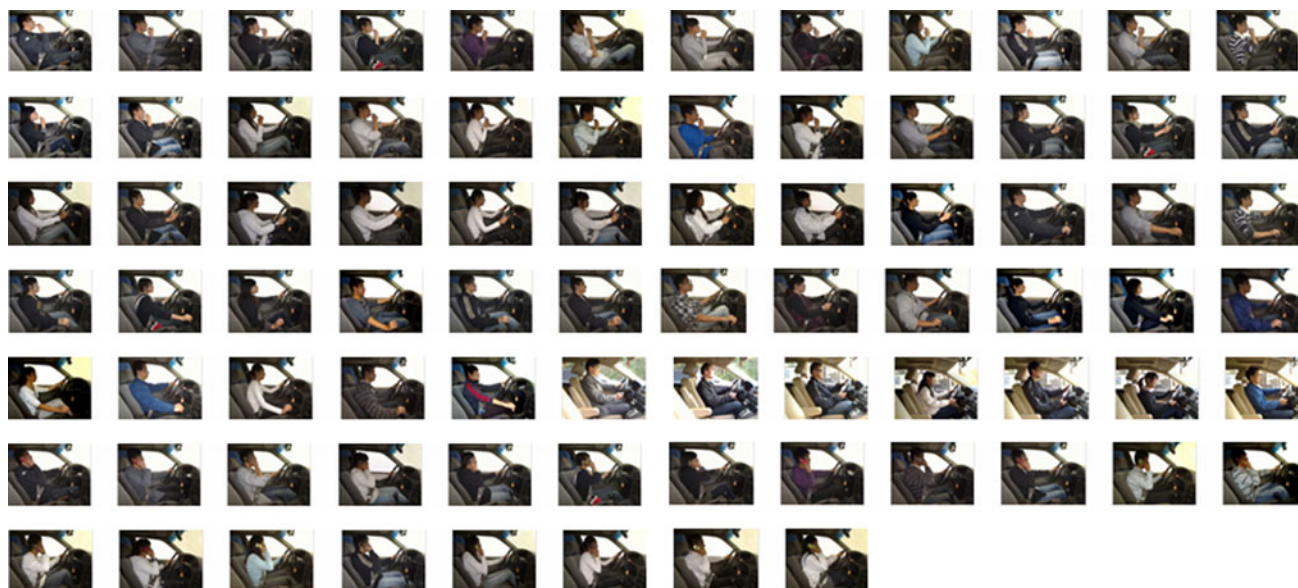


Fig. 6 Southeast University Driving Posture (SEU-DP) data set

Fig. 7 Classification performance of three feature extraction methods in holdout experiment as **a** bar plots and **b** box plots, respectively

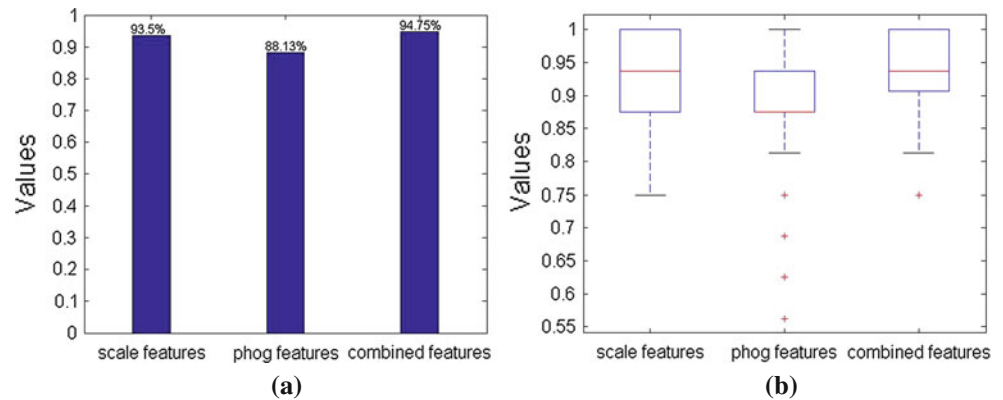


Table 1 Confusion matrix for the result from the combined feature extraction with RSE of MLP classifiers in holdout experiments

Class	I	II	III	V
I	100 %	0	0	0
II	9.65 %	90.35 %	0	0
III	0	0.96 %	99.04 %	0
V	0	0	10.84 %	89.16 %

I grasping the steering wheel, *II* operating the shift gear, *III* eating a cake, and *V* talking on a cellular phone

predicted class, respectively. The diagonal entries represent the correct classification, while the off-diagonal entries represent the incorrect ones. The rows and columns of the confusion matrices express the posture classes of grasping the steering wheel, operating the shift gear, eating a cake and talking on a cellular phone. The corresponding classification accuracies are 100, 90.35, 99.04 and 89.16 %, respectively. From the confusion matrix of the holdout experiment, among the four classes, grasping the steering wheel is the easiest to classify, while talking on a cellular phone is the most difficult with the most misclassifications.

5.2 Cross-validation experiments

The k -fold cross-validation approach is another commonly used technique that takes a set of m feature vectors and randomly partitions them into k folds of size m/k . For each fold, the classifier is tested on onefold (consists of m/k feature vectors) and trained on the other $k - 1$ folds (consisting of $m(k - 1/k)$ feature vectors). In the following experiments, 10-fold cross-validation was used when comparing histogram-based feature extraction by PHOG feature extraction, spatial scale feature extraction and the proposed combined feature extraction. The 80 feature vectors of driving postures extracted from images in SEU-DP data set were randomly divided into 10 disjoint subsets of equal size (every subset consists of 8 feature vectors of driving postures). Nine of the ten subsets were

used in training, and the remaining subset was used for testing. The experiment repeated for 10 times with each time a different testing subset.

The 80 feature vectors of driving postures extracted from images in SEU-DP data set were randomly divided into 10 folds for 100 times; 100 cross-validation experiments were carried out with each of the 100 different divisions. The histogram-based feature extraction by PHOG, spatial scale-based feature extraction and combined feature extraction were compared. The average classification accuracies of the 100 cross-validation experiments are displayed as bar plots in Fig. 8a and as box plots in Fig. 8b, respectively. The average classification accuracies of histogram-based feature extraction by PHOG, spatial scale feature extraction and combined feature extraction are 93.72, 91.16 and 94.84 %, respectively. From the bar plots and box plots of the classification rates, the proposed combined feature extraction from histogram-based features and spatial scale features by random subspace ensemble of MLP classifiers outperforms the other two feature extraction approaches.

In the cross-validation experiment, Table 2 illustrates the confusion matrix, which summarizes the detailed performance of combined feature extraction with random subspace ensemble of MLP classifiers. The classification accuracies of the four classes (i.e., grasping the steering wheel, etc.) are 99.95, 91.25, 98.60 and 89.48 %, respectively. It is shown that talking on a cellular phone is the most difficult posture to classify among the four classes, while grasping the steering wheel is the easiest to classify in the cross-validation experiment. This observation is consistent to that was obtained in holdout experiment.

6 Discussions

From Tables 1 and 2, it seems there exist some relations between the correct rate and incorrect rate for the different four classes of driving postures. For example, the class of

Fig. 8 Classification performances of the three feature extraction methods in the cross-validation experiment as **a** bar plots and **b** box plots, respectively

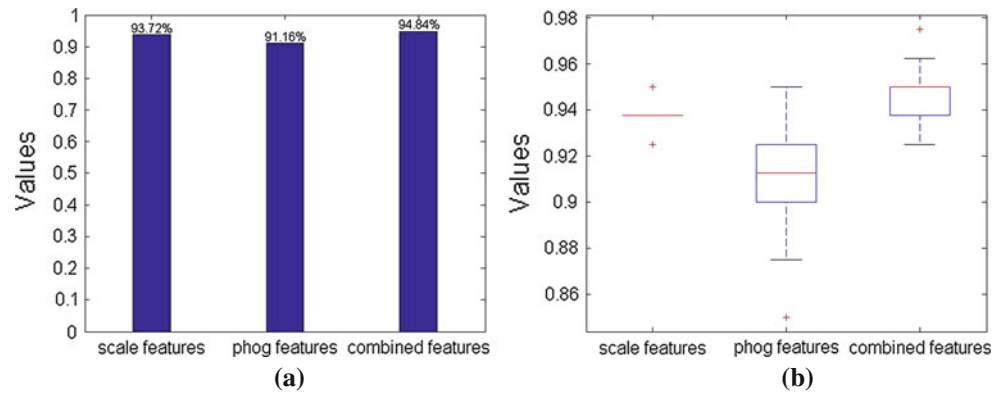


Table 2 Confusion matrix for the results from the combined feature extraction with RSE of MLP classifiers in cross-validation experiments

Class	I	II	III	V
I	99.95 %	0	00.05 %	0
II	8.60 %	91.25 %	0.15 %	0
III	0	1.40 %	98.60 %	0
V	0	0	10.52 %	89.48 %

I grasping the steering wheel, *II* operating the shift gear, *III* eating a cake, and *V* talking on a cellular phone

talking on a cellular phone will be misclassified as eating cake with an incorrect rate of 10.84 % in the holdout experiment and 10.52 % in the cross-validation experiment. But eating cake will not be misclassified as talking on a cellular phone. Instead, it will be misclassified as grasping the steering wheel with an incorrect rate of 0.96 % in the holdout experiment and 1.40 % in the cross-validation experiment. It is a very interesting observation that the misclassification is mostly one-directional. With the four postures in Tables 1 and 2, Class II, III and IV tend to be misclassified as posture I, II and III, respectively. But the reverse misclassification (e.g., Class I to Class II) did not happen.

From the experimental results, one of the possible explanations is that the postures are not symmetrically distributed in the feature space. To simplify the illustration, we consider a one-dimensional feature space below. For instance, the distributions of the four postures could be Poisson style as shown in Fig. 9. In this case, posture *d* tends to be misclassified as posture *c*. But posture *c* will not be misclassified as posture *d*, because there is no occurrence of posture *c* crossing the classification threshold between posture *c* and *d*. Similar performances can be expected in classification of posture *b* and *c*. Posture *a* will not be misclassified.

Bearing this in mind, we visually examined the images more carefully and found some clues consistent to the

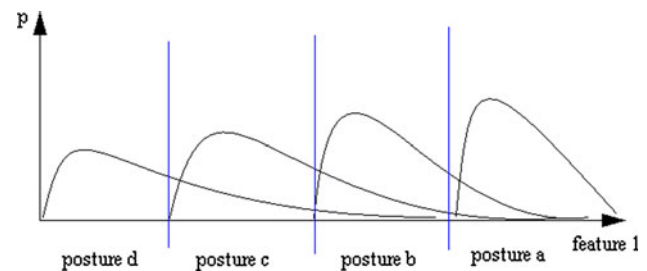


Fig. 9 Asymmetrical probability distributions in the feature space and their classifications

above explanation. Taking Class III and Class IV as examples, where Class III is eating a cake and Class IV is talking on a cellular phone. The right-hand locations generally overlap with head locations in Class IV and marginally connect to the head location in Class III. These two cases are considered mean of the distributions of the two postures. However, there are a few cases in Class IV where the cellular phones are bigger, which separate the right hand and head regions. In classification, these cases tend to be misclassified as Class III. In training, however, they can only affect a small shift of the mean but will not introduce misclassification from Class IV.

7 Conclusions

In this paper, we proposed an efficient combined feature extraction approach with random subspace ensemble of MLP classifiers, to automatically understand and characterize the driver behaviors. Four classes of driving postures were exploited in the proposed posture classification framework, which were grasping the steering wheel, operating the shift lever, eating a cake and talking on a cellular phone. The combined feature extraction was constructed based on histogram-based feature description by PHOG and spatial scale-based feature description. With feature vectors extracted from SEU-DP data set,

we comparatively studied histogram-based feature extraction by PHOG, spatial scale-based feature extraction and combined feature extraction in the classification of four predefined classes of driving postures. The holdout and cross-validation experiments indicated that the combined feature extraction offers better classification performances over two individual feature extraction approaches. Our experiments also showed that talking on a cellular phone is the most difficult posture among the four driving posture classes. With the proposed combined feature extraction and random subspace ensemble of MLP classifiers, the classification accuracies of talking on a cellular phone are over 89 % in both of the holdout and cross-validation experiments, which shows the effectiveness of the proposed combined feature extraction approach and random subspace ensemble of MLP classifiers in automatically understanding and characterizing driver behaviors toward human-centric driver assistance systems.

References

- Nadeau CL, Maag U, Bellavance F et al (2003) Wireless telephones and the risk of road crashes. *Accid Anal Prev* 35:649–660
- Oliver N, Pentland AP (2002) Graphical models for driver behavior recognition in a smartcar. In: IEEE intelligent vehicles symposium, pp 7–12
- Liu XY, Zhu D, Fujimura K (2003) Real-time pose classification for driver monitoring. In: IEEE intelligent transportation systems, pp 174–178
- Eren H, Celik U, Poyraz M (2007) Stereo vision and statistical based behavior prediction of driver. In: IEEE intelligent vehicles symposium, pp 657–662
- Sekizawa S, Inagaki S, Suzuki T et al (2007) Modeling and recognition of driving behavior based on stochastic switched ARX model. *IEEE Trans Intell Transp Syst* 8(4):593–606
- Watta P, Lakshmanan S, Hou YL (2007) Nonparametric approaches for estimating driver pose. *IEEE Trans Veh Technol* 56(4):2028–2041
- Kato T, Fujii T, Tanimoto M (2004) Detection of driver's posture in the car by using far infrared camera. In: IEEE intelligent vehicles symposium, pp 339–344
- Cheng SY, Park S, Trivedi MM (2007) Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis. *Comput Vis Image Underst* 106(2–3):245–257
- Demirdjian D, Varri C (2009) Driver poses estimation with 3D time-of-flight sensor. In: IEEE CIVVS, pp 16–22
- Cheng SY, Trivedi MM (2006) Turn-intent analysis using body pose for intelligent driver assistance. *IEEE Perv Comput* 5(4):28–37
- Yang CM, Wu CC, Chou CM et al (2009) Vehicle driver's ECG and sitting posture monitoring system. In: 9th conference on information technology and application in biotechnology, pp 1–4
- Cheng SY, Trivedi MM (2010) Vision-based infotainment user determination by hand recognition for driver assistance. *IEEE Trans Intell Transp Syst* 11(3):759–764
- Yoon HS, Soh J, Bae YJ et al (2001) Hand gesture recognition using combined features of location, angle and velocity. *Pattern Recogn* 34(7):1491–1501
- Mitra S, Acharya T (2007) Gesture recognition: a survey. *IEEE Trans Syst Man Cybern Part C Appl Rev* 37(3):311–324
- Han J, Awad G, Sutherland A (2009) Automatic skin segmentation and tracking in sign language recognition. *IET Comput Vis* 3(1):24–35
- Suk HI, Sin BK, Lee SW (2010) Hand gesture recognition based on dynamic Bayesian network framework. *Pattern Recogn* 43(9):3059–3072
- Huang DY, Hu WC, Chang SH (2011) Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination. *Expert Syst Appl* 38(5):6031–6042
- Harini V, Stefan A, Nathaniel B, et al (2005) Driver activity monitoring through supervised and unsupervised learning. In: IEEE conference on intelligent transportation systems, pp 895–900
- Veeraraghavan H, Bird N, Atev S et al (2007) Classifiers for driver activity monitoring. *Transp Res Part C* 15(1):51–67
- Heusch G, Cardinaux F, Marcel S (2005) Lighting normalization algorithms for face verification. *Tech Rep IDIAP*, pp 7–8
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition, San Diego, CA, pp 886–893
- Zhang BL (2010) Off-line signature verification and identification by pyramid histogram of oriented gradients. *Int J Intell Comput Cybern* 3(4):611–630
- Heath M, Sarkar S, Sanocki T et al (1998) Comparison of edge detectors: a methodology and initial study. *Comput Vis Image Underst* 69(1):38–54
- Hansen L, Salamon P (1990) Neural network ensembles. *IEEE Trans Pattern Anal Mach Intell* 12:993–1001
- Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. Wiley-Interscience, New Jersey
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140
- Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 55:119–139
- Boland MV, Murphy RF (2001) A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* 17:1213–1223
- Boland MV, Markey M, Murphy RF (1998) Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry* 33:366–375
- Huang K, Murphy RF (2004) Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics* 5:78
- Haykin S (2000) An introduction to neural networks: a comprehensive foundation, 2nd edn. Prentice-Hall, Upper Saddle River
- Skurichina M, Robert P, Duin W (2002) Bagging, boosting and the random subspace method for linear classifiers. *Pattern Anal Appl* 5:121–135
- Ho TK (1998) The random subspace method for constructing decision forest. *IEEE Trans Pattern Anal Mach Intell* 20:832–844