Driving Posture Recognition by a Hierarchal Classification System with Multiple Features

Chao Yan, Bailing Zhang
Department of Computer Science & Software Eng.
Xi'an Jiaotong-Liverpool University
Suzhou, 215123, China

Frans Coenen
Department of Computer Science
The University of Liverpool
Liverpool, UK

Abstract—**This paper presents a novel system for vision-based driving posture recognition. The driving posture dataset was prepared by a side-mounted camera looking at a driver's left profile. After pre-processing for illumination variations, eight action classes of constitutive components of the driving activities were segmented, including normal driving, operating a cell phone, eating and smoking. A global grid-based representation for the action sequence was emphasized, which featured two consecutive steps. Step 1 generates a motion descriptive shape based on a motion frequency image(MFI), and step 2 applies the pyramid histogram of oriented gradients (PHOG) for more discriminating characterization. A three level hierarchal classification system is designed to overcome the difficulties of some overlapping classes. Four commonly applied classifiers, including k-nearest neighbor(KNN), random forest (RF), support vector machine(SVM) and multiple layer perceptron (MLP), are evaluated in each level. The overall classification accuracy is over 87.2% for the eight classes of driving actions by the proposed classification system.**

Keywords-component; driving posture recognition; driving assistance system; motion frequency image; hierarchal classification

I. INTRODUCTION

Unsafe and dangerous driving accounts for the death of more than one million lives and over 50 million serious injuries worldwide each year [1]. The WHO also estimates that traffic accidents cost the Chinese economy over $21 billion each year. One of key contributing factors is reckless driving. An emerging technology that has attracted worldwide attention is the development of an intelligent driver assistance system that continuously monitors not just the surrounding environment and vehicle state, but also the driver's behaviors. One of the proposed solutions for automatic understanding and characterization of unsafe driving behaviors such as fatigue, eating and talking on a cellular telephone is by using a camera-based system to monitor the activities of drivers.

Previous works of vision-based driving activities monitoring mainly focused on the eyes, face, head, facial expressions or an appropriate combination [2-4]. Oliver and Pentland [5] proposed a machine learning framework for modeling and recognizing a driver's movements, with emphasizes on how the context affects the driver's performance, using graphical models, hidden markov models (HMMs) and coupled hidden markov models (CHMMS). Kato et al. proposed a system with a far-infrared camera to minimize the influence of variations from illumination and [6]. Cheng et al. presented a multi-camera method [7]. Veeraraghavan et al. and Zhao et al. recognized driving actions by exploiting drivers' skin-region information [8-9]. Tran et al. studied driver's behavior by foot gesture analysis [10].

The task of driver behavior monitoring can be generally studied in the human action recognition framework, the emphasis of which is on finding good feature representations that could be robust to the variations in viewpoint, human subject, background, illumination and so on. This paper extends our previous work[11] which has yielded satisfying results by the joint application of motion history image and pyramid histogram of oriented gradients on driving action period recognition. In this paper, we studied the drivers' activity recognition by comprehensively considering illumination variation pre-processing, action frame detection, action clips segmentation, action clips representation and hierarchal classification. Our contributions include three parts. The first part is our deviation from many published works on drivers' posture based on static images, which has the potential problem of confusion caused by similar postures. As a result, we regard driving activity as space-time action instead of static space-limited posture and proposed a method for action segmentation from the original video data. The second contribution is the proposal of a normalization algorithm to reduce the influence from illumination variation caused by the camera's built-in intensity compensation function. The motion detection and the further action segmentation all highly depend on the accuracy of frame difference, which is very sensitive to illumination variation. The last contribution of this paper is the proposal of the application of a hierarchal classification system for driving posture recognition. Some subsets of classes are difficult to classify due to their overlapping in some feature subspace. By treating original multi-classes as a hierarchal classification problem, overlapping problems can be alleviated.

The rest of the paper is organized as follows. Section 2 gives a brief introduction on the SEU driving posture dataset creation and the overview of the recognition system. Section 3 explains the proposed method for illumination variation pre-processing. Section 4 introduces the action segmentation algorithm. Section 5 reviews the concept of motion frequency image, with explanations of how they are applied in driving posture description. Section 6 gives the details of the hierarchal classification system. Section 7 reports the experiment results, followed by conclusion in Section 8.

II. DATASET ACQUISITION AND SYSTEM OVERVIEW

To test the proposed driving posture recognition approach, we used the SEU driving dataset first created by Zhao [9]. The dataset is recorded by a side-mounted Logitech C905 CCD camera under day lighting conditions. Ten male drivers and ten female drivers participated in the creation of the dataset by pretending to drive in the car and conducting several activities that simulated real driving situations. Five pre-defined driving activities were imitated, that is, turning the steering wheel, operating the shift lever, eating, smoking, and using a cell phone.

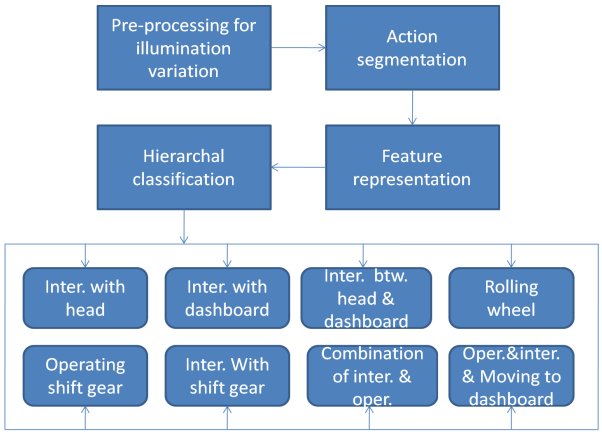

Figure 1. System overview

As shown in Fig.1, our proposed driving posture recognition system consists of four main steps, that is, pre-processing for illumination variations, action segmentation, feature representation, and hierarchal classification. There are eight classes of driving actions in the segmented sequence. Class 1 is head interaction, where the driver takes something towards or away from his/her head by using his/her right hand. For example, moving food towards the mouth or taking a call by bringing the cell phone towards the ear. Class 2 is the interaction with the car dashboard; the driver takes something or puts something back on to the dashboard. For example, taking a cigarette or putting back the cigarette lighter. Class 3 is a combination of class 1 and class 2, including some interactive hand motions between the head and dashboard. Classes 4 and 5 refer to the normal manipulation of the steering wheel and shift gear, respectively. Class 6 of the action includes the interactive hand motions between right hand and gear shift. Class 7 is a compositional action constituted by action 5 and action 6. The last class of action is defined for the situation where right hand is first operating the gears, then moving back to the steering wheel and then reaching to the dashboard.

III. PRE-PROCESSING FOR ILLUMINATION VARIATIONS

Most cameras have a built-in intensity compensation function in soft level which increases the global intensity if it is low and vice versa. It aims to increase the contrast of the subject in the camera. But the responding time of this function is too slow when increasing or decreasing the global intensity. Therefore, when the function works as the camera is set in recording video mode, it will make the intensity of frame sequence unstable. As a result, it has a negative influence on movement detection by using frame difference from a gray-level image. Fig.2(b) gives an example of a frame difference. The white area represents the moving point but there is no movement between frame 3334 and 3335 actually. It is caused the camera built-in intensity compensation function. Fig.2(e) shows the intensity value versus frame number in video 25. There are many leaps and the intensity value stays stable for a few frames in different levels of intensity value. These leaps are caused by the intensity compensation function of the camera. Theoretically, this negative influence can be removed by reversing the algorithm of the built-in intensity compensation function.

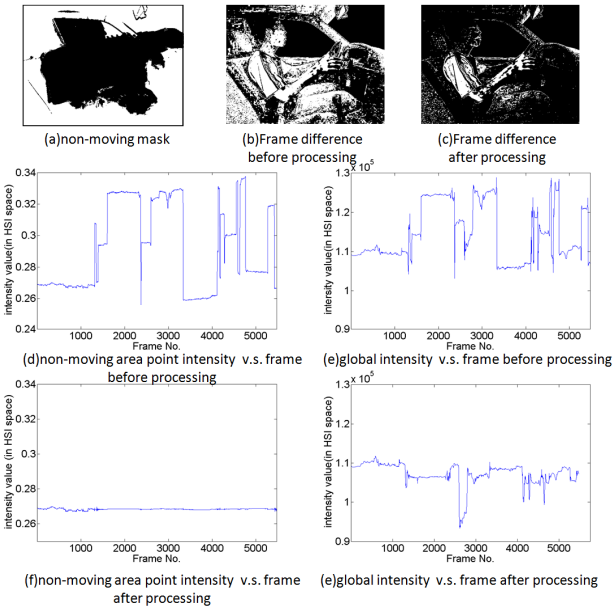

Figure 2. Reduction in intensity variation caused by intensity compensation

In our work, the negative influence is reduced as follows.Fig.2(a) is a non-moving area mask of video clip 25. The white area represents the non-moving area in the steering room. It is extracted by stacking the frame difference images from each pair of consecutive frames in sequence. The frame difference image is thresholded by Otsu's method [12]. Fig.2(d) shows the point intensity value of non-moving area versus frame number. The point intensity value is stable around 0.27 in previous frames, for about 1400 frames. The reason is that the driver did not perform any activities except slightly turning the steering wheel in the first 1400 frames. The content in the first 1400 frames is almost static, so the global intensity is stable and the intensity compensation function doesn't work. The average intensity of each point on non-moving area of the first 1400 frames is defined to be the point intensity constant (PIC) in an assumption of neglected natural illumination variation. The formal definition of PIC can be expressed by:

$$\mathrm{PIC} = \frac{\sum_1^W \sum_1^L \sum_1^{1400} (\mathrm{im_I} \times \mathrm{Mask_{non\_moving}})}{\mathrm{Area}(\mathrm{Mask_{non\_moving}}) \times 1400} \qquad (1)$$

where W is image width, L is image length, $\mathrm{im_I}$ is image intensity, $\mathrm{Mask_{non\_moving}}$ is a binarymask forthe non-moving area among the frame sequence and $\mathrm{Area}(\mathrm{Mask_{non\_moving}})$ is

the total point number of the non-moving area. To reduce the influence of intensity compensation for a given image in the sequence, the function is given by

$$\mathrm{im}_I'' = \mathrm{im}_I' + \mathrm{diff}_I \tag{2}$$

$$\mathrm{diff}_I = \frac{\sum_1^W \sum_1^L (\mathrm{im}_I' \times \mathrm{Mask}_{\mathrm{non\_moving}})}{\mathrm{Area}(\mathrm{Mask}_{\mathrm{non\_moving}})} - \mathrm{PIC} \tag{3}$$

whereim_I'isthe intensity of a given image on which we need to reduce the influence from intensity compensation function.$\mathrm{diff}_I$is the difference intensity that needs to be applied on the given image intensity. $\mathrm{im}_I''$is the intensity adjusted image fromim_I'. Fig.2(f) is the point intensity value of non-moving area versus frame number after applying the above method. The point intensity value is stable around 0.27. Fig.2(g) is the image intensity value versus frame number after applying the method. It is much more stable compared to the plot in Fig.2(e). Fig.2(b) is the original frame difference image of frame 3335 in video 25. Fig.2(c) is the corresponding frame difference image after the application of the above method. The white area in frame difference image which represents movement is much less than the original one after applying above algorithm.

IV. DRIVING ACTION SEGMENTATION

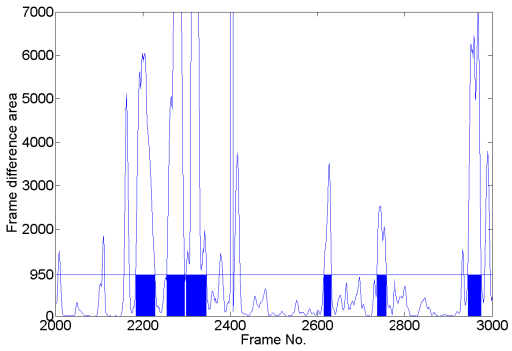

Figure 3. Action segmentation

In the original SEU dataset, each video represents one participant driving in the steering room. The driver performs activities including operating the shift gear, responding to a cell phone call, eating and smoking while driving. There is neither segmented action period nor classification standard in the original dataset. How to segment the raw driving action sequence is not a trivial problem. In our works, video sequence segmentation is based on the continuity of at least 15 frames with which frame difference area is over 950 points.

Fig.3 further explains practice by plotting the frame difference area versus frame number, which shows that six action clips are segmented between frames 2000 to 3000 from video 25 of the SEU dataset. By using this method, 527 action clips are segmented out from all 20 videos of SEU dataset. The threshold values of frame difference area and period length are empirical values chosen from the experiment. For different datasets, the threshold value may vary due to different background noise, action period precision and other practical factors.

V. MOTION FREQUENCY IMAGE (MFI)

Action clips segmented from the original video is a sequence of high-dimensional images, which cannot be directly applied for classification. To simplify the problem, we take an average of the image sequence to generate a Motion Frequency Image(MFI) as a representation of the original image sequence, as shown in Fig.4.

Figure 4. Illustration of extracting MFI

MFI is a view-based temporal template approach inspired by Motion History Image (MHI) which describes how the motion is moving in the image sequence [13]. Given the binary frame difference images $\mathrm{B_t}(\mathrm{x},\mathrm{y})$at time t in a sequence, the gray-level motion frequency image (MFI) is defined as follows:

$$\mathrm{MFI}(\mathrm{x},\mathrm{y}) = \frac{1}{\mathrm{N}}\sum_{\mathrm{t}=1}^{\mathrm{N}} \mathrm{B_t}(\mathrm{x},\mathrm{y}) \tag{4}$$

where N is the number of frames, t is the frame number in the sequence, and x and y are values in the 2D image coordinate. MFI, also named Gait Energy Image (GEI) in individual gait recognition [14], indicates the movement frequency and phase information. Another related representation is the motion energy image (MEI) [13], which demonstrates the presence of any motion or a spatial pattern in the image sequence.

VI. HIERARCHAL CLASSIFICATION SYSTEM

The MFIs of segmented action clips from the SEU dataset are labeled into eight classes as described above. Some subsets of classes are difficult to classify due to the overlap in some feature subspace. The multi-class problem can be best treated in a hierarchal classification framework. Many researchers have concluded that hierarchal classification architectures outperform flat classification techniques [15-17]. As shown in Fig.5, we first classify the segmented periods into shift gear related and shift gear non-related class, each of which will be then further classified in a lower level of the hierarchal classification system as some subsets of the classes are closer to other subsets. Different regions of interest(ROI) and low-level features can be exploited for the different subsets in the proposed top-down hierarchal classification system.

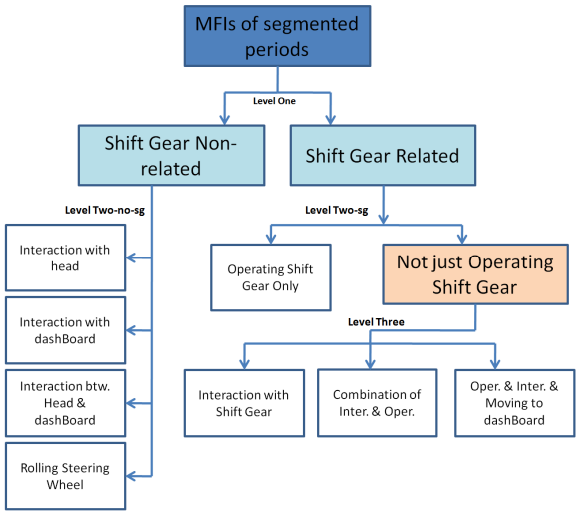

Figure 5. Hierarchal classification system

A. *Level one classification*

The first level consists of a SVM classifier[18-19], which is trained to make a distinction between the shift gear related and shift gear non-related actions. The driving actions taken place in the steering room is mainly performed by hand which motivates the definition of these two classes. As shown in Fig.6, it is a skin region time lapse image. When the driver is operating the shift gear or interacting with the shift gear, the arm and hand will appear in the right lower box, the shift gear related area and vice versa. As a result, in this first level of classification, the region of interest (ROI) is set to be the shift gear related area. The motion energy images (MEI) in the ROI of the two classes are shown in Fig.7 which demonstrates a significant difference.

B. *Level two classification*

There are two branches in the 2nd level of class hierarchy. The first branch (abbreviated as level two-sg in the figure) categorizes two situations, namely, *operating shift gear only* and *not only operating shift gear*. A random forest classifier [20] is trained to classify the two groups of pattern as shown in Fig.8. The second branch (abbreviated as level two-no-sg in the figure) covers the following four cases: interaction with head, interaction with dashboard, hand moving between head and dashboard, and normal manipulation of the wheel, as shown in Fig.9. Similar to the previous discussion, random forest classifier is trained to classify the four groups of MFI.

C. *Level three classification*

In the third level of classification hierarchy, three subclasses of the *not only operating shift gear* class are defined, as shown in Fig.10. They are close in MFI feature space and difficult to classify. As these three actions are performed by the right hand mainly moving among shift gear, steering wheel and dashboard, the trajectories of the right hand are easily distinguishable. One possible approach to locate the right hand is by using skin-region analysis in a well-defined region of interest (ROI). We further extract the right hand skin-region in a ROI for each image of the action sequence, and combine them to form a right hand skin-region MFI. There are many methods

Figure 6. ROI based on skin region time lapse image

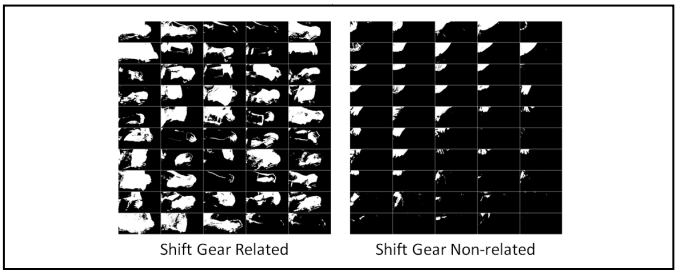

Figure 7. Two classes in level one of the hierarachal classification system

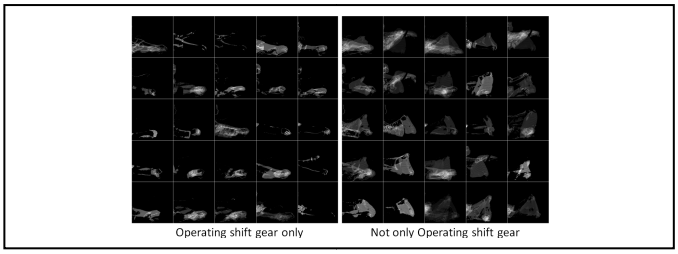

Figure 8. Two classes in level two-sg of the hierarachal classification system

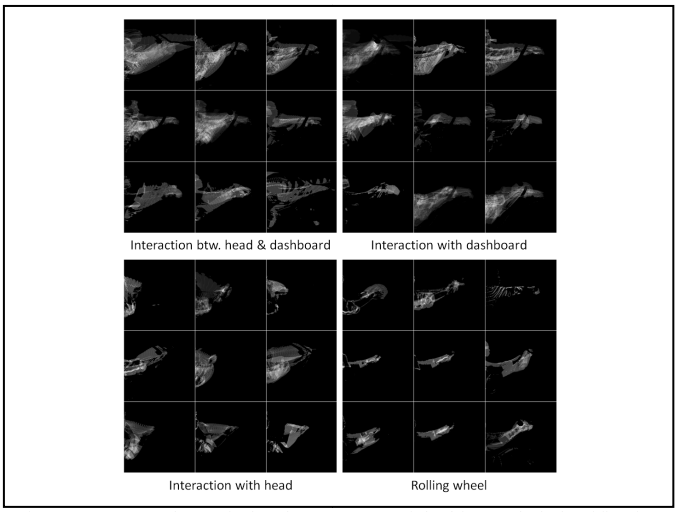

Figure 9. Four classes in level two--no-sg of the hierarachal classification system

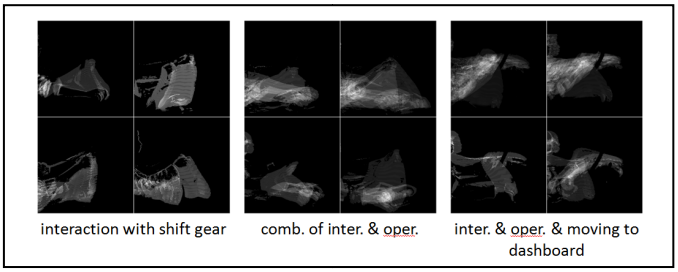

Figure 10. Three classes in level three of the hierarachal classification system

for skin region segmentation such as difference color space thresholding, Gaussian and mixture of Gaussian distributions thresholding method[21]. In this experiment, we simply segment the region of skin if the value in YCbCr color space lies within the following rules.

$$\begin{cases} 80 \leq \mathrm{Cb} \leq 120 \\ 140 \leq \mathrm{Cr} \leq 170 \end{cases} \quad (5)$$

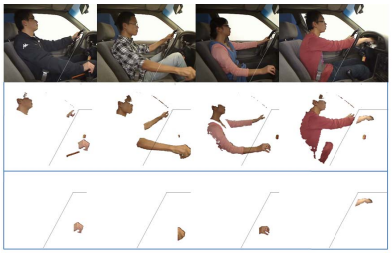

Figure 11. Locating the right hand skin region in ROI

Fig.11 shows the procedure of locating the right hand skin region in ROI. The first row is four selected frames from the original sequence. The second row is the skin region after applying above rule corresponding to the first row. It works well on Asian skin, but with the problem that skin-tone clothing may cause confusion. As the three classes of action are related to the shift gear region and the dashboard region, the region of interest(ROI) is located at a right trapezoid region of the lower right corner of the frame, which covers the shift gear region and the dashboard region. We only estimate the right hand region in ROI. The third row shows the hand region in ROI after connected component analysis and further analysis of the hand area.

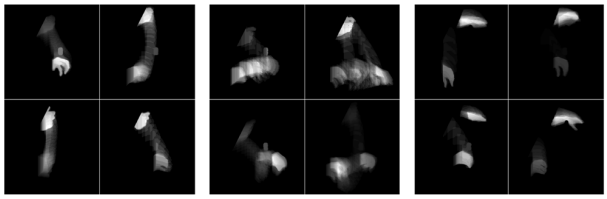

Figure 12. Three classes in level three of the hierarachal classification system after right hand skin-region extraction

After locating the right hand skin region in ROI for each frame in the sequence, the right hand region sequence is combined to form another group of MFI as shown in Fig.12, which is much easier to classify compared to the pattern in Fig.10.

VII. EXPERIMENT

In the experiment, 20 videos from the original SEU dataset are first pre-processed to reduce the influence of illumination variation. After that, the action clips are segmented by the algorithm discussed in section IV. Then eight different classes of sequences are sent to the hierarchal classification system for training and classification. We chose a standard experimental procedure called the holdout approach to verify the driving action performance using KNN, RF, SVM and MLP classifier.

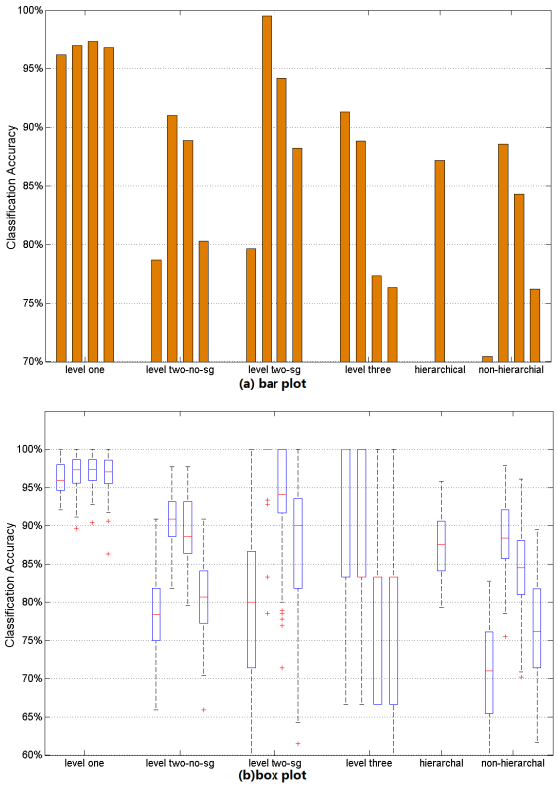

Figure 13. Experiment result of bar plot and box plot

In the holdout experiment, 10% of the 20 videos, that is 2 videos, are randomly selected as the testing dataset, while the remaining 18 videos are used as the training dataset. The average results from 100 runs are shown in Figure 13(a) and the classification accuracy distribution results in 100 runs are shown in Figure 13(b). The Table I is the numerical results of the bar plot in Figure13(a). Based on the performance shown in Table I, we choose SVM in level one, RF in level two and KNN in level three to form the hierarchal classification system, and the final classification accuracy is 87.2%. It is 1.37% lower than the non-hierarchal classification result of 88.75% which only applies MFI and PHOG in a one-versus-eight RF classifier.

TABLE 1CLASSIFICATION ACCURACY

	Classification Accuracy (%)			
	KNN	***RF***	***SVM***	***MLP***
Level one	96.35	96.98	97.12	96.96
Level two-no-sg	78.68	91.02	88.86	80.32
Level two-sg	79.63	99.48	94.18	88.20
Level three	91.33	88.83	77.33	76.33
Hierarchal systme	87.2			
No hierarchal	70.47	88.57	84.31	76.22

However, the confusion matrix in Fig.14 proves the significance of the hierarchal system. In confusion matrix of

non-hierarchal system, the accuracy of action 5 is only 19%,which mistakes for action 2 with the rate of 59% and action 7 with the rate of 23%. The accuracy of action 5 is increased to 79% in the confusion matrix of hierarchal system which means that 81% subsets of action 5 is closer to the others class in the feature space of combination of MFI and PHOG.

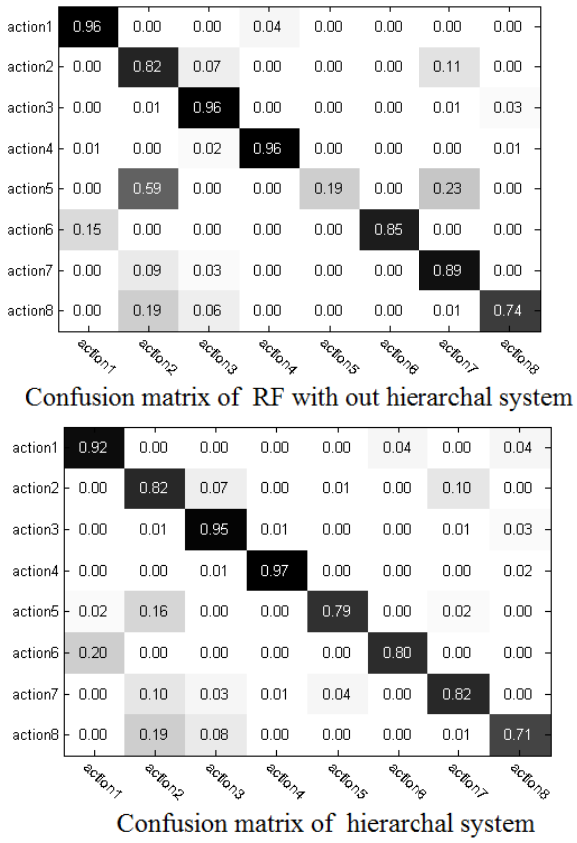

Figure 14. Confusion matrix

VIII. CONCLUSTION

This paper addresses the importance of automatic understanding and characterization of driver behaviors in preventing motor vehicle accidents and presents a novel system for vision-based driving posture recognition. We test our approach on the SEU driving posture dataset which includes activities of normal driving, responding to a cell phone, eating and smoking. After pre-processing for illumination variations and action clip segmentation, eight classes of actions are extracted for classification. By joint application of motion frequency image, pyramid histogram of oriented gradients, hand skin-region segmentation and the hierarchal classification, our overall accuracy is over 87.2%. While the overall accuracy decreases 1.37% compared to non-hierarchal classification system, the individual classification accuracy for each class increases to no less than 71%.

REFERENCES

[1] "WHO World report on road traffic injury prevention," http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/en/

[2] P. Watta, S. Lakshmanan, Y. Hou, Nonparametric approaches for estimating driver pose. IEEE Transactions on Vehicular Technology, Vol.56, no.4, pp.2028-2041,2007

[3] E. Wahlstrom, O. Masoud, N.Papanikolopoulos, Vision-based methods for driver monitoring. in Proceedings of IEEE Intelligent Transportation Systems, Vol.2, pp.903-908, Shanghai, China, October 2003.

[4] A. Doshi, M. M. Trivedi, On the Roles of Eye Gaze and Head Dynamics in Predicting Driver's Intent to Change Lanes. IEEE Transactions on Intelligent Transportation Systems, Vol.10(3), pp.453-462, 2009.

[5] Oliver, N., Pentland, A.P.: Graphical models for driver behavior recognition in a smart car. Proceedings of IEEE Intelligent Vehicles Symposium, pp. 7–12, 2002.

[6] T. Kato,T. Fuji, M. Tanimoto, Detection of drivers posture in the car by using far infrared camera. Proceedings of IEEE Intelligent Vehicles Symposium, pp.339-344, Parma, Italy, 2004.

[7] S. Y. Cheng , S.Park, M. M. Trivedi, Multi-spectral and multi-perspective video arrays for driver body tracking and activity analysis. Computer Vision & Image Understanding, Vol.106(2-3), pp.245-257, 2007.

[8] H.Veeraraghavan, N. Bird, S. Atev, N. Papanikolopoulos, Classifiers for driver activity monitoring. Transportation Research Part C: Emerging Technologies, Vol.15, (1), pp. 51-67, 2007.

[9] C.Zhao, B.Zhang, J. He, J. Lian, Recognition of driving postures by contourlet transform and random forests. IET Intelligent Transport Systems, Vol.6 (2), pp.161-168, 2012.

[10] C. Tran, A. Doshi, M. M. Trivedi, Modeling and prediction of driver behavior by foot gesture analysis. Computer Vision and Image Understanding, Vol.116(3), pp. 435-445, 2012.

[11] C. Yan, F. Coenen, and B. Zhang,Driving Posture Recognition by Joint Application of Motion History Image and Pyramid Histogram of Oriented Gradients,International Journal of Vehicular Technology, vol. 2014, Article ID 719413, 2014.

[12] N. Otsu, A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, Vol.9(1), pp.62-66,1979.

[13] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.23(3), pp.257-267, 2001.

[14] J. han, B. Bhanu, Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.28(2), pp.316-322, 2006.

[15] J. X. Dong, A. Krzyzak, and C.Y. Suen, Fast SVM training algorithm with decomposition on very large datasets. IEEE Transactions Pattern Analysis and Machine Intelligence, Vol.27(4), pp.603–618, 2005.

[16] H. Sahbi, D. Geman, A Hierarchy of Support Vector Machines for Pattern Detection. Journal of Machine Learning Research, Vol. 7, pp.2087–2123, 2006.

[17] C. N. Silla Jr., A. A. Freit, A survey of hierarchal classification across different application domains. Data Min Knowl Disc. , Vol.22, pp.31–72, 2011

[18] V. Kecman, Learning and Soft Computing, Support Vector machines, Neural Networks and Fuzzy Logic Models, The MIT Press, Cambridge, MA, 2001.

[19] L.P. Wang (Ed.), Support Vector Machines: Theory and Application, Springer, Berlin, 2005.

[20] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[21] W. R. Tan, C. S. Chan, P. Yogarajah, and J. Condell, A fusion approachfor efficient human skin detection, IEEE Transaction Industrial Informatics, vol.8, no. 1, pp. 138–147, 2012