Driver Behavior Recognition Based on Deep Convolutional Neural Networks

Shiyang Yan*, Yuxuan Teng*, Jeremy S.Smith† and Bailing Zhang*
*Department of Computer Science and Software Engineering
Xi'an Jiaotong-Liverpool University
SIP, Suzhou, China, 215123
†Department of Electrical Engineering and Electronics
University of Liverpool
Liverpool, L69 3BX, UK

***Abstract*—Traffic safety is a severe problem around the world. Many road accidents are normally related with the driver's unsafe driving behavior, e.g. eating while driving. In this work, we propose a vision-based solution to recognize the driver's behavior based on convolutional neural networks. Specifically, given an image, skin-like regions are extracted by Gaussian Mixture Model, which are passed to a deep convolutional neural networks model, namely R*CNN, to generate action labels. The skin-like regions are able to provide abundant semantic information with sufficient discriminative capability. Also, R*CNN is able to select the most informative regions from candidates to facilitate the final action recognition. We tested the proposed methods on Southeast University Driving-posture Dataset and achieve mean Average Precision(mAP) of 97.76% on the dataset which prove the proposed method is effective in drivers's action recognition.**

***Keywords*:—Driver's Behavior Recognition; Skin-color Modeling; Gaussian Mixture Model; Convolutional Neural Networks; R*CNN.**

I. Introduction

Improvement of road safety is one of the most important objectives for Intelligent Transportation, which has become ever-increasingly more urgent with the expected increase in vehicle numbers in the future. Most of the traffic accidents are, at least in part, due to human error. A driver's behavior plays a critical role in this circumstance. The monitoring system in an intelligent vehicle would able to provide accurate information indicating if a driver's action is allowed. We aim to find a vision-based solution of the driver action recognition in this paper. To be more specific, convolutional neural networks are applied for the action recognition on still image dataset built based on real driving condition.

Most of the previously published works on action recognition were on video-based approaches [1] [2] [3]. While sequential movement information is often characteristics to some actions, many actions can still be recognized from still images without motion information. This has attracted attentions in recent years [4] [5] [6] [7]. The action can be reliably identified based on the actor's pose and the object. Action recognition from still images is not only important to many real-world problems, but also fundamental to the action recognition in video.

Like general image classification, previous research on action recognition mainly depends on certain kind of hand-designed features, for example, LBP [8], HoG [9] or SIFT [10], which are then applied by appropriate machine learning algorithms [11] [12] for different tasks. The hand-designed features are not optimized for visual representation and is generally separated stage in the overall system. Accordingly, efficient and discriminative feature expression is the main bottleneck in most of the previous vision problems, including action recognition. In recent years, an emerging machine learning area, i.e., deep learning, has attracted wide attention with the objective for feature learning and extraction. And there has ever-increasing evidence that deep learning in general, and a special kind of deep learning models, namely, Convolutional Neural Networks (CNNs) [13] [14] [15] [16], is the most promising way to learning visual representations. There have been many works [17] [18] [19] on the application of deep convolutional neural networks on action recognition from still image. Among them, R*CNN proposed by Gkioxari et al. [17] is the one of most influential ones. The common approach for action recognition in still image explores contextual cues such as human pose and human-object interactions [6] [20] [21]. In the basic framework of R*CNN, the process of context discovery is automatically achieved in convolutional neural networks.

In this paper, we aim to detect driver's behavior within the deep learning framework by exploring both the pose information of the driver and the contextual cues of the image. Specifically, we follows the framework of R*CNN [17] with certain improvements. We explore the contextual cues which contribute the most for certain actions by exploring the skin-like region. Skin regions are firstly extracted by Gaussian Mixture Model (GMM) algorithm [22] trained by skin images, then forwarded to R*CNN [17] to discriminate a driver's action.

We summarize the contributions of this paper as follows:

(1) We improve the conventional R*CNN framework by replacing generic region proposal algorithms with skin-like region extractor.

(2) Our method provides a vision-based solution for a driv-

978-1-5090-4093-3/16/$31.00 ©2016 IEEE

er's action recognition in the field of Intelligent Transportation.

(3) The final results of the proposed system achieved satisfactory results on Southeast University Driving-posture Dataset (SEU dataset).

II. System overview

Figure. 1 illustrates the structure of our system. Aiming to classify a driver's action, we first recreated a groundtruth data with driver's action defined and labelled. Secondly, the full image is processed by a GMM which has been trained with skin samples. We manually collected skin images sampled from different human skin types including white, brown, black and yellow. The GMM will generate skin-like regions which is considered as the secondary region in R*CNN framework [17]. Secondary region, originally generated by selective search [23] in conventional R*CNN, is replaced with skin regions which are more informative. Lastly, R*CNN jointly process the primary region and secondary region for finally generating action class labels.

A. Brief Introduction of Gaussian Mixture Model

For skin color distribution modeling, there are generally two approaches: parametric such as Gaussian Mixture Model (GMM) [22] and non-parametric such as histogram of Bayesian probability [24] [25]. In this paper, a parametric approach, i.e., GMM mothod, is applied.

For more robust skin color distribution modeling, during both training and testing, the color images are firstly transformed to YCbCr color space. In conventional RGB color space, a pixel color is related with all three variables, namely Red, Green and Blue. This simple approach is not able to sperate color, luminance and chrominance. The YCbCr color space can represent luminance and chrominance in sperate variables by Y which stands for luminance while the other two represents for chrominance.

Once obtained YCbCr images, Gaussian model can be applied on modeling skin color. Although gaussian models have been successfully applied to represent features of images in various practical problems, the assumption of single entity in gaussian distribution means a too simplistic distribution of skin color which does not match with reality and can cause intolerable error.

$$P(x|skin) = \frac{1}{\Pi|\Sigma|^{-1/2}} exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \tag{1}$$

Equation (1) is the density function of Gaussian distribution where μ is the expectation vector and Σ is the covariance matrix of the normally distributed variable x. A better solution is to represent the distribution in reality by combining values generated from different sources [26]. GMM can be considered as weighted sum of single gaussian distribution [26].

$$P(x) = \sum_{n=1}^{k} w_i \frac{1}{\Pi|\Sigma|^{-1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)\right\} \tag{2}$$

Equation (2) is the density function of GMM. k is the total number of Gaussian components of the model and w_i is the i^{th} weight of the corresponding components.

B. Brief Introduction of R*CNN

Recently, along with the excellent performance of deep convolutional neural networks in object detection and image classification [16] [15], many works have been published to implement convolutional neural networks in action recognition [18] [17]. Among them, R*CNN, proposed in [17], make use of all available contextual cues for final action recognition. These contextual cues, are either the object that interact with person or specific human parts that contribute mostly for a actor's action.

According to [17], the inputting regions of R*CNN includes two parts, the primary region and secondary region. Primary region, i.e., the ground truth, provide person in question. And the secondary region, which is all the bounding boxes generated by a region proposal algorithm, provides contextual cues for action recognition. The primary region and secondary region are forwarded to R*CNN for jointly training action recognition. The processing of primary region is similar with a conventional deep convolutional neural network. Yet, R*CNN brings a simple mechanism that utilize max operation on classification scores of secondary region, which achieves automatic selection of the most influential contextual parts for the final recognition. This scheme, can not only discover contextual cues but also be applicable to fine-grained recognition as well as attributes discovery [17]. Then, the classification scores of primary region and secondary region are added as inputs to Softmax layer for final outputs. The system structure of R*CNN is illustrated in Figure. 2.

In the original R*CNN framework, the region proposal is generated by selective search [23]. However, as for a driver's behavior recognition, intuitively, the skin-like regions are the most possible part that could relate to a driver's behavior. As a result, instead of using generic region proposal algorithm, we apply GMM for skin color modeling. The skin-like regions can be generated firstly, and inputted into R*CNN for selecting the top contributor for action recognition.

III. Experiment

A. Dataset Introduction

To test the proposed approach on driver behavior recognition, the Southeast University Driving-posture Dataset(SEU dataset) was used. The dataset was built by Zhao [27]. Twenty drivers participated in the establishment of the dataset [27]. We extracted key frames from this video dataset, and randomly select images for training and testing separately. Also, we manually annotated person region(ground truth) and corresponding classes for the dataset which include 6 categories of behavior:

(1) Call: Responding to phone call.
(2) Eat: Eating while driving.
(3) Brake: operating the shift gear.
(4) Wheel: correct driving position with hands on wheel.

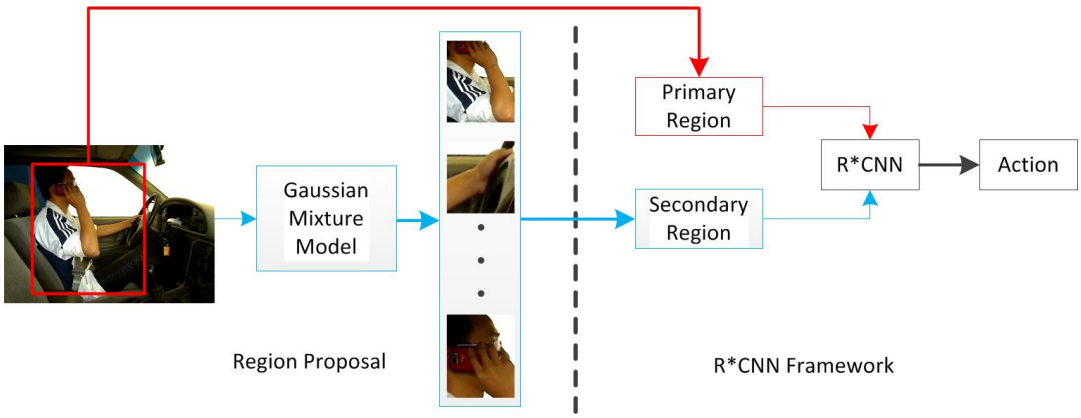

Fig. 1: System Overview

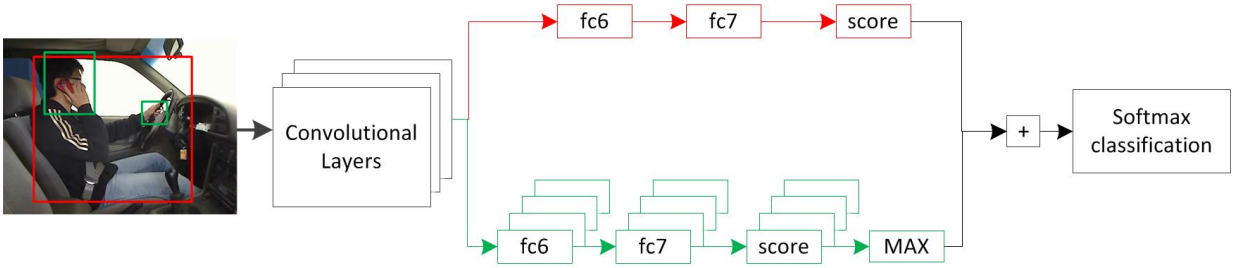

Fig. 2: Principle of R*CNN

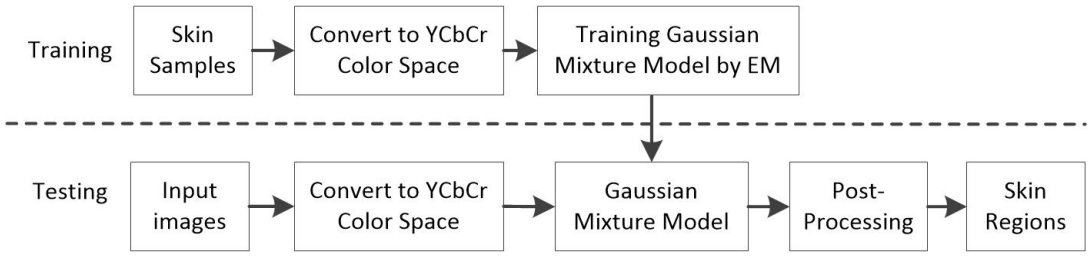

Fig. 3: Implementation of Gaussian Mixture Model for Skin Modeling

Fig. 4: Examples of SEU dataset

(5) Phone-play: Playing phone while driving.

(6) Smoke: Driving while smoking.

Figure. 4 provides some examples of SEU dataset in which each image from left to right corresponds with different behavior following the sequence from category (1) to (6).

Chao Yan et al. [28] also exploited this dataset for driver posture recognition based on convolutional neural networks and achieved satisfactory results. However, our target task is different from [28] as we define 6 action types including very similar pairs in appearance, i.e., "Eat" and "Smoke", contrasting with only 4 driving activities in [28].

B. Implementation Platforms

Our experiment was conducted on a dell Tower 5810 with Intel Xeon E5-1650 v3 and memory 64GBs. In order to speed up CNN training, a GPU, NVIDIA GTX TITAN, is plugged on the Workstation. The program is operated on the 64-bit Open-source Linux operating system CentOS 7 installed with CUDA 7.5, Python2.7.3 and Matlab 2014b, also Caffe deep learning platform is used for CNN training and testing.

Specifically, for GMM training and testing, Matlab is utilized. The region proposals generated by GMM are forward-

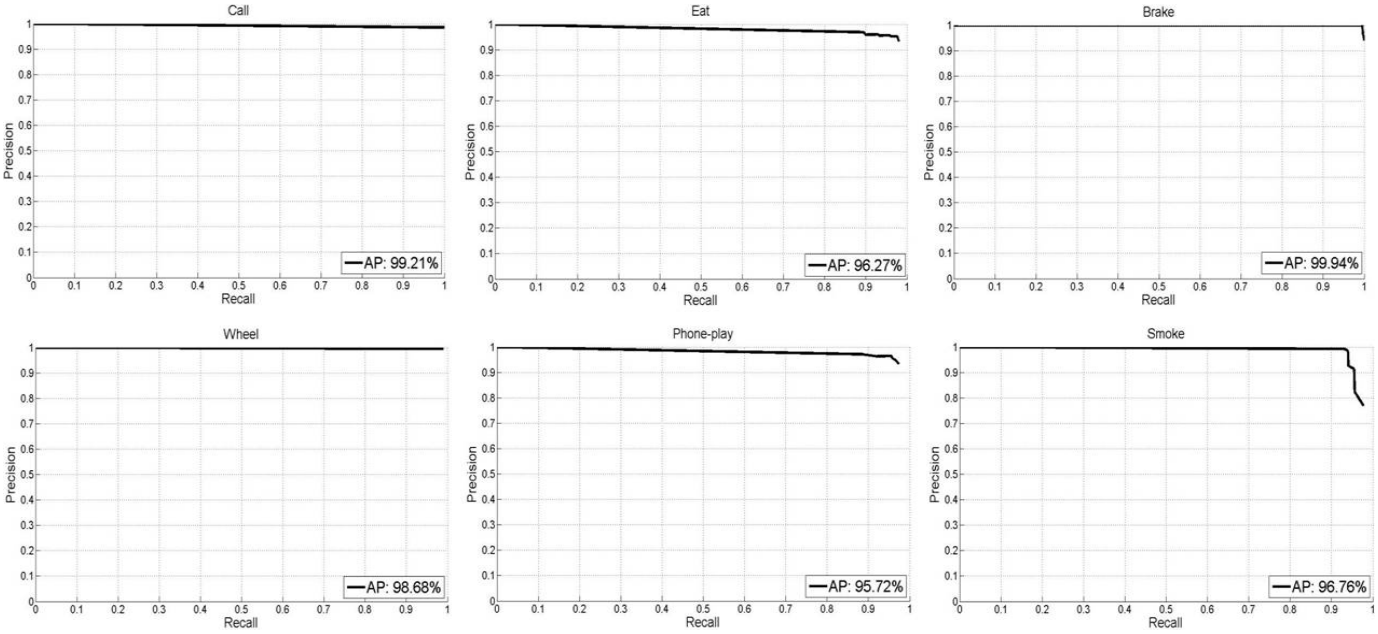

Fig. 5: Results of the System

Fig. 6: Visualization of Secondary Regions

ed to R*CNN for final recognition. Following the common practice of pre-training CNN [15], the parameters of layers of CNN-M from [29] is assigned. Then the CNN model is fine-tuned with ground truth region and candidate regions from GMM.

C. Implementation Details of Gaussian Mixture Model

Figure. 3 clearly illustrates the procedure of how to implement GMM. The skin samples are firstly converted to YCbCr color space for subsequent GMM training. The GMM is trained via standard Expectation Maximization(EM) method. In testing, similarly, images are transformed to YCbCr space and then modeled by GMM, the output feature map are post-processed by binarization and dilation to generate final skin-like regions.

D. Experimental Results of the System

To evaluate the system performance, following the standard evaluation procedure of action recognition [17], the AP is calculated for each action category. First of all, the precision-recall curve (PR curve) is plotted for each class as shown in Figure. 5. The area under PR curve, i.e., AP, is computed accordingly. The highest AP is achieved on the class of "Brake", which implies this action is easier to discriminate because the pose of a hand perform braking is obviously different from other poses. As for the action class of "Eat" and "Smoke", the poses are extremely similar only with differences on the objects that interact with hands. Even though, the final results are satisfactory which further prove the applicability of our system for more fine-grained tasks. To further evaluate our system performance, the mean AP of 97.76 is finally computed. In addition, we also implement three traditional approaches using hand-crafted features on this dataset. TABLE. I clearly

TABLE I: Final Results of Different Methods

AP(%)	Call	Brake	Wheel	Eat	Phone-play	Smoke	mAP
Our method	99.21	96.27	99.94	98.68	95.72	96.76	97.76
DSIFT+MLP	92.50	93.32	95.34	67.57	80.75	60.69	81.67
PHOG+MLP	93.25	89.91	92.35	57.86	90.98	61.39	80.96
LBP+MLP	61.80	72.83	73.17	44.86	74.24	53.57	63.41

shows that our method largely outperform traditional ones. To be more specific, for schemes based on DSIFT [30], PHOG [31] and LBP [8] features with Multi-layer Perceptron(MLP) classifier, the AP results on "Eat" and "Smoke" are the poorest, which in turn prove the superiority of the our methods on fine-grained action discrimination.

E. Visualization of Selected Secondary Region

For more straightforward illustration, Figure. 6 provides some examples of the drivers in question and activated secondary region corresponding to a certain action class. Each column corresponds to different action. Red boxes highlight the driver to be classified while blue boxes automatically selected the secondary region. As can be seen from Figure. 6, secondary region normally implies the regions containing cues of hand pose. This further implies the effectiveness of the skin-like region proposal and powerful discriminative capability of R*CNN.

IV. Conclusion

Aiming to recognize a driver's behavior inside vehicle, this paper provides a vision-based system based on deep convolutional neural networks. Following the basic framework of R*CNN but with skin-like region extractor (GMM) replacing generic region proposal e.g.Selective Search, along with the location providing a driver in question, our proposed scheme is able to successfully discriminate each action type, even when two actions are extremely similar. Furthermore, the satisfactory results on Southeast University Driving-posture Dataset prove the applicability of the proposed approach in driver's behavior recognition. Future works will be conducted on more efficient region proposal for R*CNN and improvement of the present CNN structure.

References

[1] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[2] F. Husain, B. Dellen, and C. Torras. Action recognition based on efficient deep feature learning in the spatio-temporal domain. *IEEE Robotics and Automation Letters*, 1(2):984–991, July 2016.

[3] Chaur-Heh Hsieh, Mao-Hsiung Hung, and Wei-Yang Huang. Recognizing human actions by fusing spatial-temporal hog and key point histogram. *Journal of the Chinese Institute of Engineers*, pages 1–10, 2016.

[4] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1331–1338. IEEE, 2011.

[5] Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC 2010-21st British Machine Vision Conference*, 2010.

[6] Bangpeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1691–1703, 2012.

[7] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361, 2014.

[8] Dong-Chen He and Li Wang. Texture unit, texture spectrum and texture analysis. In *Geoscience and Remote Sensing Symposium, 1989. IGARSS'89. 12th Canadian Symposium on Remote Sensing., 1989 International*, volume 5, pages 2769–2772. IEEE, 1989.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[10] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[11] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[12] Xunshi Yan and Yupin Luo. Making full use of spatial-temporal interest points: an adaboost approach for action recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4677–4680. IEEE, 2010.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–97. IEEE, 2004.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[17] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r*cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[18] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2470–2478, 2015.

[19] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. R-cnns for pose estimation and action detection. *arXiv preprint arXiv:1406.5212*, 2014.

[20] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 9–16. IEEE, 2010.

[21] Gang Yu, Zicheng Liu, and Junsong Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Computer Vision–ACCV 2014*, pages 50–65. Springer, 2014.

[22] Ming-Hsuan Yang and Narendra Ahuja. Gaussian mixture model for human skin color and its applications in image and video databases. In *Electronic Imaging'99*, pages 458–466. International Society for Optics and Photonics, 1998.

[23] Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. Segmentation as selective search for object recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1879–1886. IEEE, 2011.

[24] Michael J Jones and James M Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[25] Chengjun Liu. A bayesian discriminating features method for face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(6):725–740, 2003.

[26] Jie Yang and Alex Waibel. A real-time face tracker. In *Applications of Computer Vision, 1996. WACV'96., Proceedings 3rd IEEE Workshop on*, pages 142–147. IEEE, 1996.

[27] C. H. Zhao, B. L. Zhang, J. He, and J. Lian. Recognition of driving postures by contourlet transform and random forests. *IET Intelligent Transport Systems*, 6(2):161–168, June 2012.

[28] C. Yan, F. Coenen, and B. Zhang. Driving posture recognition by convolutional neural networks. *IET Computer Vision*, 10(2):103–114, 2016.

[29] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[30] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, May 2011.

[31] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.