DETECTING EMOTIONAL STRESS FROM FACIAL EXPRESSIONS FOR DRIVING SAFETY

Hua Gao, Anil Yüce, Jean-Philippe Thiran

Signal Processing Laboratory (LTS5),
École Polytechnique Fédérale de Lausanne, Switzerland

ABSTRACT

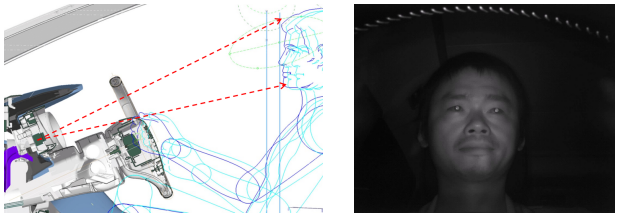
Monitoring the attentive and emotional status of the driver is critical for the safety and comfort of driving. In this work a real-time non-intrusive monitoring system is developed, which detects the emotional states of the driver by analyzing facial expressions. The system considers two negative basic emotions, anger and disgust, as stress related emotions. We detect an individual emotion in each video frame and the decision on the stress level is made on sequence level. Experimental results show that the developed system operates very well on simulated data even with generic models. An additional pose normalization step reduces the impact of pose mismatch due to camera setup and pose variation, and hence improves the detection accuracy further.

Index Terms— emotion, stress, detection, driver, face, facial expression

1. INTRODUCTION

Applying modern computer vision technologies for enhancing safety of vehicle driving has been investigated for several decades. Most of the research focused on detecting drowsiness of the driver, which according to [1], causes a large percentage of the car accidents. Recently, reports [2, 3] also show that the emotional status (e.g. stress, impatience) of the driver may as well endanger the safety. From the viewpoint of behavior scientists, high level stress may damage self-confidence, narrow attention and eventually disrupt concentration. This often leads to aggressive driving and makes the driver pay less attention to the traffic situation. To reduce riskiness from a stressed state, it is necessary to detect such emotions and take certain actions to relax the driver.

Most of the previous work on stress detection applies physiological features (such as electromyogram, electrocardiogram, respiration, and skin conductance) [4, 5]. It is found that in real-world driving tasks, skin conductivity and heart rate metrics are most closely correlated with driver stress level [5]. However, those measurements are intrusive, so are less comfortable in real applications. A non-intrusive stress detection system is developed in [6], in which a physiological measure based on skin temperature is used. In [2], acoustic signals are used for measuring the stress level. However, the performance might be affected by the noisy in-car driving environment. A system in [7] fuses several physiological signals and visual features (eye closure, head movement) to monitor driver drowsiness and stress in a driving simulator. Liao et al. [8] applies facial expression, head motion and eye gaze as the visual cues for stress inference and evidences of the different signal modalities are combined with Dynamic Bayesian Networks (DBN). [9] classifies the driver's facial emotion from thermal images, which provide a natural combination of visual evidence and skin temperature.



(a) (b)

Fig. 1: Camera setup: (a) An NIR-camera is mounted inside the dashboard and directed towards the driver's face; (b) An example of captured face image.

In this work, we developed a stress detection system based on the analysis of facial expressions. The system is non-intrusive and is able to run in real-time, which allows immediate reaction upon detection. A near-infrared (NIR) camera is used to capture the near frontal view of the driver's face. The camera is mounted inside the dashboard facing the driver. A face tracker is applied to track a set of facial landmarks. The holistic or local texture features are extracted to classify facial expression classes in each frame. To compensate for the pose mismatch due to head motion and camera view, a pose normalization step is applied which generates a virtual frontal view of the face. The inferences of expressions on the frame level are integrated and a decision on stress detection is made for a moving time window on sequence level. We evaluate the proposed system on two recorded datasets. Experimental results demonstrate that the system has a detection rate of 90.5% for the in-door tests and 85% for the in-car tests.

2. SYSTEM OVERVIEW

In this section, we briefly describe the hardware and software setup of the developed system. Figure 1(a) sketches the camera configuration in this system. An NIR camera is mounted inside the dashboard behind the steering wheel of

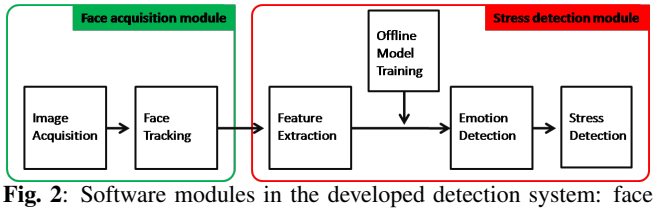


Fig. 2: Software modules in the developed detection system: face acquisition module and stress detection module.

a car, with a slightly up-tilted view-angle towards the driver. The NIR-camera is used to compensate for the illumination effects caused by ambient lighting conditions. Figure 1(b) shows an example of the captured face image. Note that the face might be partially occluded by the steering wheel when the driver turns it.

The acquired NIR-video frames are processed sequentially by the software modules as illustrated in Figure 2. The face acquisition module detects and tracks the driver's face in real time. This module also provides a set of facial landmarks for the subsequent stress detection module. In the stress detection module, holistic or local texture features are extracted from the normalized facial images. Classifiers, which are trained offline with the extracted features, are applied to determine the presence of facial expressions. A temporal fusion is employed to integrate decisions of individual frames and expression classifiers, which eventually determines the level of stress according to the facial appearance of the driver.

3. EMOTIONAL STRESS DETECTION

This section describes the implementation details of the individual modules. This includes the definition of the detection target and the methodologies. We also discuss potential problems due to the camera setup.

3.1. Task Definition

Different people may behave or express differently under stress. It is hard to find a universal pattern to define the stress emotion. Moreover, it is also not easy to collect data for training offline models. To ease the problem, we define stress in terms of the basic emotions, which have corresponding facial expressions that are well defined and universal [10]. We define the expression of stress as anger, disgust, or a combination of these two fundamental facial expressions. In this preliminary study, we consider that stress is detected if either anger or disgust is detected constantly within a fixed time interval.

3.2. Methodologies

3.2.1. Face Tracking

The face acquisition module processes the video sequences captured by the NIR-camera. It detects and tracks the driver's face in real-time. In addition to the face position, this module also provides a set of facial landmarks, which depict the geometry and temporal motion of individual facial components. Alternatively, face pose orientation can also be estimated with the tracked landmarks.

We apply the supervised descent method (SDM) [11] as our face tracker. Basically, the method assumes that a valid shape can be estimated with a cascade of regression models, given an initial guess of the shape. The regression models are learned stage-wise with local texture features $\mathbf{\Phi}_t$ extracted around the current estimated landmarks, with respect to their estimation residues $\mathbf{r}_t$, i.e. $\mathbf{r}_t = \mathbf{R}_t * \mathbf{\Phi}_t$. The matrix $\mathbf{R}_t$ stands for the regression model at stage t. Note that the images are normalized to a standard scale, and a fixed scale of local texture patches (e.g. SIFT [12] descriptor) are extracted.

The SDM face tracker requires an initialization step. The well known Viola & Jones face detector [13] is applied for initialization, which provides a bounding box indicating the location and scale of the detected face. The tracker fits a shape in the initial frame starting with the mean shape and continues the fitting in the succeeding frames. Figure 3 shows an example frame of the face tracking results overlaid with the 49 tracked facial landmarks.

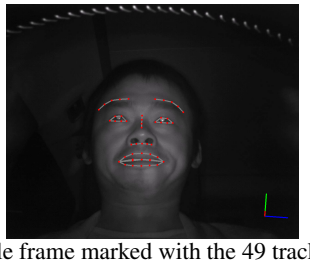


Fig. 3: A sample frame marked with the 49 tracked landmarks.

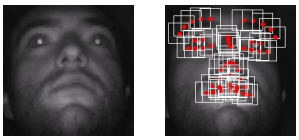
3.2.2. Feature Extraction

Several approaches are applied to extract discriminative features to learn patterns of different facial expressions. We investigate approaches based on holistic affine warping and local descriptors.

The holistic affine warping method normalizes face images using the coordinates of the left eye and right eye. The locations of the eye centers are derived from the tracked facial landmarks. After applying an affine transform, the eye centers are fixed in the canonical coordinates in the normalized image. Figure 4(a) shows an example of the normalized face image with holistic affine warping. The local DCT feature representation [14] is extracted from the normalized image, in which low frequency DCT features are obtained in non-overlapping local blocks.

The second approach extracts local descriptors directly around the tracked facial landmarks. It preserves the geometrical information of the facial components and does not introduce additional artifacts. Following [15], we extract SIFT descriptor centered at each landmark. The orientation and scaling parameters of the SIFT descriptors are fixed. Each facial image is normalized to a standard scale: 200×200 pixels. The SIFT descriptor is extracted in a 32×32 local block. Figure 4(b) illustrates an example of the local descriptor-based

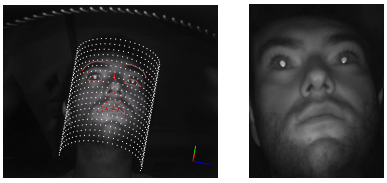
method. All the extracted SIFT descriptors are concatenated, and we apply PCA for dimensionality reduction. The final feature vector preserves 98% variance of the original data. A similar approach is also applied in applications such as face identification [16].



(a) (b)

Fig. 4: Approaches of feature extraction: (a) Holistic affine warping; (b) Local descriptors: extracting SIFT descriptors from local patches centered at the tracked facial landmarks.

Due to pose variation or camera setup, the face pose in the test frames may not match the training data very well. To mitigate the impact of pose mismatch on detection performance, we adopt a pose correction method. Based on the tracked facial landmarks, we estimate head pose parameters with the least squares method. A 3D Cylindrical Head Model (CHM) [17] is applied with a simplified 3D face surface. Using the estimated head pose parameters we rotate the CHM (see Figure 5(a)) and project 2D texture pixels onto the CHM with bi-linear interpolation. Figure 5(b) shows a pose normalized face image. The tracked facial landmarks are mapped on the CHM and again the local DCT representation or SIFT descriptors can be extracted.



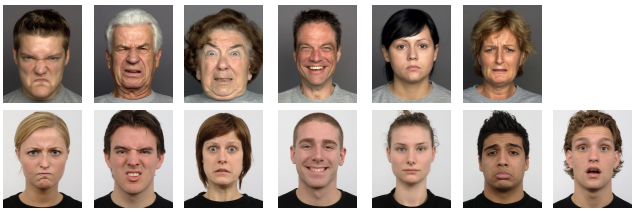
(a) (b)

Fig. 5: Pose normalization with CHM. (a) A 3D CHM projected on a 2D image; (b) Texture mapped on CHM with bi-linear interpolation.

3.2.3. Detection

As defined in Section 3.1, the stress detection is based on detecting basic facial expressions. We consider Ekman's six basic emotions [10] and neutral expression as in most works on facial expression recognition in the literature, e.g. [18]. The six basic emotions are anger, disgust, fear, happiness, sadness and surprise. To train classifiers for recognition, we collect images from two facial expression databases, i.e. the FACES [19] database and the Radboud [20] database. The first row in Figure 6 shows example images from the FACES database, in which the expression of surprise is not present. The second row shows examples from the Radboud database. The FACES database contains 179 subjects while the Radboud database contains 49 subjects. The images in both databases are in frontal and upright pose, and evenly illuminated with posed facial emotions.

We train multi-class classifiers with the extracted feature representations. The classifiers are implemented with support vector machines (SVM) in a one-vs-all manner. Linear-SVMs are trained and the parameters are tuned with a five-fold cross validation. For testing, we consider the distance to hyperplane as the classification confidence of each individual binary classifier. The class with the highest confidence is considered as our classification hypothesis in each individual frame. To make the estimation more robust against noise in face tracking, we set a moving time window W_1 with F_1 frames. The decision is determined with a voting of expressions representing the stress, i.e. anger and disgust, against the rest of the emotions within the time window. In addition, another moving time window W_2 (with F_2 frames) is used to determine the degree of stress, i.e. what percentage of frames is classified as stress in W_2. If the percentage exceeds a threshold τ_1, we consider the driver is under stress.



anger disgust fear happiness neutral sadness surprise

Fig. 6: Sample images used for training emotion classifiers. 1st row: the FACES database; 2nd row: the Radboud database.

Data set	#subjects	#videos	recording condition
Set1	21	42	frontal view, office setting
Set2	12	20	up-tilted view, car setting

Table 1: Recorded data sets for evaluating stress detection.

4. EXPERIMENTS AND ANALYSIS

To evaluate the performance of the developed system, we collected data and carried out quantitative experimental evaluation. The results are discussed and justified in this section.

4.1. Data Collection

Two data sets were recorded to evaluate stress detection. Set1 was recorded in an office with an NIR-camera placed on a desk in front of the recorded subjects. None of the recorded subjects are professional actors. The subjects were asked to pose the six basic emotions and the neutral expression. The recorded frames were used for model adaption in the later experiments. For each subject we recorded two additional videos for testing. Each recording captures two minutes of frames at around 25 fps with a resolution of 1280×1024. The subjects were asked to pose the expression of stress for one minute starting at 30 seconds after the beginning. In total 21 subjects were recorded in Set1.

Set2 was recorded in a car, where the NIR-camera was placed exactly as illustrated in Figure 1(a). This data set simulates the real application scenario. The recording protocol was the same as in Set1. Twelve subjects were recorded in Set2, five of which also appear in Set1. Table 1 summarizes the details about Set1 and Set2.

4.2. Experimental Results

The results of the stress detection experiments are presented in Table 2. The second column in Table 2 lists the results obtained on Set1 and the third column lists the results achieved on Set2. Results of the proposed feature extraction approaches are compared. HAW denotes the holistic affine warping method and LD denotes the approach based on local descriptors. Pose normalization (PN) using CHM is also applied with both methods, which correspond to PNHAW and PNLD. The postfix "+" indicates that a model adaptation is applied with additional training data recorded in Section 4.1. For each subject, five frames are used for adapting each emotion classifier. The adaption is done by building new classifiers which are trained using original training images and the collected frames. Note that no test frames are used in the offline model training. The results denote the number of successfully detected videos depicting stress (out of 42 videos on Set1 and 20 videos on Set2). We also present the frame level detection results in terms of F-measure in the parentheses. All the results are achieved with a common parameter setting, i.e. $F_1 = 10$, $F_2 = 450$, and $\tau_1 = 2/3$.

In general, the local descriptors method performs better than HAW on Set1. The HAW-based method detects one video more than the LD-based method, but the F-measure is decreased, which means that the frame-wise detection precision of the HAW-based method is lower than the LD-based method. The pose normalization step is more helpful for the LD-based method. We also observe that the adaptation step is very critical, which yields a significant performance boost. We believe, and here prove, that the model adaptation is necessary, as the facial appearance in an NIR-frame is very different from a normal facial image which captures the visible spectrum of light. With adaptation and pose normalization, the LD-based method (PNLD+) detects 90.5% of the videos in Set1, with the highest F-measure of 0.871 (recall: 0.860, precision: 0.882). A similar observation is made for Set2, in which the camera is mounted slightly up-tilted in a car for the real application scenario. The detection performance is degraded due to the change in camera setting. In particular the LD-based method performs worse than the simple HAW-based method without adaptation and pose normalization. With adaptation applied, the detection rate of LD+ is the same as for the HAW+, whereas LD+ achieves a higher F-measure. The PNLD+, which applies pose normalization in addition, achieves a detection rate of 85% (17 out of 20 videos) with the highest F-measure of 0.815 (recall: 0.735, precision: 0.914) on Set2. Figure 7 shows the detected frames (indicated with green bars) in an example video from Set2. The red crosses indicate the labeled frames in which stress is present. The stress level exceeds the threshold τ_1 about 20 seconds after the subject starts posing stress.

To demonstrate the generalizability of the developed system, we tested our system using LD+ with a camera setting similar to Set1. In total 136 subjects were tested and 83.09% of the test videos were correctly detected. Note that the expression models are only adapted to the subjects in Set1. This result indicates that an adaptation to camera setting with images of a few subjects is sufficient for a subject independent system.

Data set	Set1 (42 videos)	Set2 (20 videos)
HAW	16 (0.537)	7 (0.279)
HAW+	22 (0.732)	16 (0.737)
LD	15 (0.563)	3 (0.173)
LD+	34 (0.838)	16 (0.770)
PNHAW	15 (0.553)	8 (0.331)
PNHAW+	28 (0.790)	14 (0.741)
PNLD	18 (0.555)	8 (0.422)
PNLD+	**38** (**0.871**)	**17** (**0.815**)

Table 2: Evaluation of stress detection on Set1 and Set2. The results correspond to the number of successfully detected videos using different methods, and the frame based F-measures in parentheses.

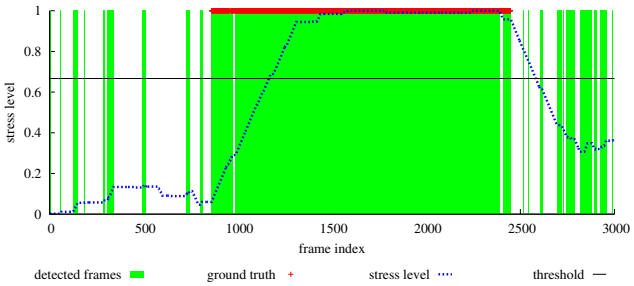


Fig. 7: Stress detection in an example video in Set2.

5. CONCLUSION AND FUTURE WORK

We developed a monitoring system for detecting emotional stress of the vehicle driver. To assess the detection performance, we conducted experiments on two collected data sets. Quantitative results on stress detection show good performance with local descriptor-based feature representation, when additional data is collected for model adaptation. A pose normalization step is applied to further mitigate the impact of pose mismatch due to camera setting and pose variation. The proposed best system is able to successfully detect 90.5% of in-door and 85% of in-car test cases.

We believe it is necessary to investigate further the definition and subject-dependent characteristics of emotional stress. A more sophisticated online adaptation can be provided which allows a user to train a person-specific stress detection model. The pattern of temporal dynamics of facial expressions and actions can also be integrated in the model training. Other cues such as head motion and acoustic signals could also be integrated to achieve better performance.

6. ACKNOWLEDGEMENT

The authors would like to thank Olivier Pajot and Estelle Chin from Stellab - PSA Peugeot Citroën for their valuable contributions in the conceptual design.

7. REFERENCES

[1] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 596–614, 2011.

[2] C. Nass, I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion," in *Intl. Conf. on HCI*, 2005.

[3] C. L. Lisetti and F. Nasoz, "Affective Intelligent Car Interfaces with Emotion Recognition," in *11th Intl. Conf. on HCI*, 2005.

[4] J. Healy and R. Picard, "Smartcar: detecting driver stress," in *Joint 3rd Int'l Workshop on Nonlinear Dynamics and Synchronization (INDS)*, 2000.

[5] J. Healy and R. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," *IEEE Transactions on Intelegent Transportation Systems*, vol. 6, no. 3, 2005.

[6] H. Kataoka, H. Yoshida, A. Saijo, M. Yasuda, and M. Osumi, "Development of a skin temperature measuring system for non-contact stress evaluation," in *the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1998, vol. 2, pp. 940–943.

[7] M. Rimini-Doering, D. Manstetten, T. Altmueller, U. Ladstaetter, and M. Mahler, "Monitoring driver drowsiness and stress in a driving simulator," in *First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2001, pp. 58–63.

[8] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, "A real-time human stress monitoring system using dynamic bayesian network," in *IEEE Conference on Computer Vision and Pattern Recognition - Workshops*, 2005, pp. 70–70.

[9] A. Kolli, A. Fasih, F. Al Machot, and K. Kyamakya, "Non-intrusive car driver's emotion recognition using thermal camera," in *Joint 3rd Int'l Workshop on Nonlinear Dynamics and Synchronization (INDS)*, 2011, pp. 1–5.

[10] P. Ekman, "Universals and Cultural Differences in Facial Expression of Emotion," *J. Cole ed. Nebraska Symposium on Motivation*, vol. 19, pp. 207–282, 1972.

[11] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539.

[12] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] P. Viola and M. J. Jones, "Robust Real-Time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[14] H. K. Ekenel and R. Stiefelhagen, "Local Appearance based Face Recognition Using Discrete Cosine Transform," in 13^{th} *European Signal Processing Conference (EUSIPCO 2005)*, Antalya, Turkey, 2005.

[15] W. S. Chu, F. De la Torre, and J. F. Cohn, "Selective Transfer Machine for Personalized Facial Action Unit Detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3515–3522.

[16] M. Everingham, J. Sivic, and A. Zisserman, "Hello My name is Buffy Automatic naming of characters in TV video," in *In BMVC*, 2006.

[17] J. Xiao, T. Moriyama, T. Kanade, and J. Cohn, "Robust Full-Motion Recovery of Head by Dynamic Templates and Re-registration Techniques," *International Journal of Imaging Systems and Technology*, vol. 13, pp. 85–94, September 2003.

[18] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 94–101.

[19] N. C. Ebner, M. Riedinger, and U. Lindenberger, "FACES–a database of facial expression in young, middle-aged, and older women and men: development and validation," *Behavior Research Methods*, vol. 42, no. 1, pp. 351–362, 2010.

[20] O. Langner, R. Dotsch, G. Bijlstra, and D. H. J. Wigboldus, "Presentation and validation of the Radboud Faces Database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.