Drivers Stress Identification in Real-World Driving Tasks

Saira Bano
Department of Information Engineering – University of Pisa (Italy)
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Pisa (Italy)
Email: saira.bano@phd.unipi.it
Supervised by: Dr. Nicola Tonellotto, Dr. Alberto Gotta

Abstract In the past few years, cross-modal distillation has garnered a lot of interest due to the rapid growth of multi-modal data. In this paper, we study stress recognition of the drivers corresponding to the driving situation. Our method enables us to recognize stress from unlabeled videos. We perform cross-modal distillation based on wearable physiological sensors and videos from on-board cameras. In this cross-modal distillation, knowledge is transferred from sensor to vision modality.

***Index Terms*—Cross-modal Transfer, Stress Detection, Deep Learning**

I. Introduction

In recent years, human emotion recognition has attracted a lot of interest in a variety of industries, including healthcare, human-machine interaction, smart homes, and automotive [1]. In this paper, we develop the methodology to detect the driver's stress in real-world driving situations. According to international research driver stress and frustration is considered as one of the important aspects of intelligent transportation systems (ITS) as a large number of road accidents occur due to drivers being distracted or stressed [2]. Driver's stress and frustration can be increased due to several reasons, such as, increase in traffic density, sharp turns, driving in mountain areas, and the complexity of the traffic environment. Moreover, distracting effects of engaging in non-driving tasks can also influence driving. Driver stress and emotion recognition has been studied, using physiological signals, facial expression from videos, and gesture [3]. There exists a significant link between a human's emotional state and their physiological reaction as emotional classification module can receive a signal from a sensor, and this biological signal has a natural emotional state related to the autonomic nervous system's control. However, in recent years, as deep learning techniques have advanced, image and video-based approaches for emotion recognition have become more effective in real-world applications [4]. Furthermore, multiple modalities include different information and explain different elements of emotions. Therefore, integrating this information from the various modalities can be more attractive to construct a robust emotion recognition model.

In this paper, we present a method to measure the driver's stress using physiological sensors and videos captured from the in-vehicle camera. Physiological sensors will help to provide the ground truth and feedback about the driver's state to the unlabeled videos. Our work is based on the hypothesis that there exists a strong co-relation between sensor and vision data as the human emotion shown by sensors is correlated with the expressions shown in videos. To exploit this common knowledge in two modalities we use the approach of cross-modal distillation to show that sensor information can be transferred to the visual domain to detect emotions, particularly stress in our case.

The subsequent sections are organized as follows. Section II gives an introduction to the state of the art. In section III a detailed description of cross-modal distillation is explained. In Section IV we provide the conclusion of this work.

978-1-6654-3595-6/21/$31.00 ©2021 IEEE

II. Background and state of the art

A. Cross Modal Distillation

Knowledge Distillation is an effective training strategy used for model compression proposed by Hinton [5], in which they transferred knowledge from a pre-trained ensemble (teacher) model to a small and lightweight student model, suitable for deployment on resource-constrained devices. This knowledge transfer between teacher and student can be achieved in many different ways such as by matching output logits of student and teacher model [6], or by minimizing the cross-entropy loss between teacher and student model outputs [7]. However, in this knowledge distillation process, knowledge is transferred from teacher to student within the same modality while in cross-modal distillation, knowledge is transferred among models belonging to different modalities. Since the advent of this cross-modal distillation, a lot of researchers have looked into this idea by analyzing the correspondence between two different modalities in teacher-student style structures. Albanie et al. exploit the relationship between unlabeled video and audio data for emotion recognition in the speech data [8]. In [9], authors demonstrated the process of transferring supervision from labeled RGB images to unlabeled depth and optical flow images. However, in this work, we are using the correspondence between sensor-vision data to detect the driver's stress.

III. Cross Modal Distillation

In this section, we describe the procedure of cross-modal distillation using two different modalities of sensor and vision. The main goal of this work is to build effective representations for the stress identification in the drivers without access to

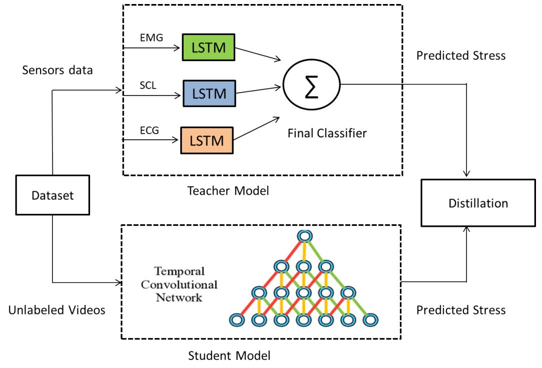

Figure 1. Proposed Methodology for Cross-modal Distillation

labeled videos. The teacher and student models are trained on physiological sensor data and unlabeled videos respectively. The key notion of this cross-modal approach is that the student model (trained on one modality) learns to mimic the features of the teacher model that is trained on different modalities and for which labels are available. Importantly, the paired inputs are considered to have the same properties in terms of the task at hand.

1) The Teacher: In order to construct a strong teacher model, rather than using one big network, we use an ensemble of LSTM modules that are trained on different sensors and outputs separate predictions for each sensor modality. Then we use a final classifier to accumulate the prediction of each of these learning modules to make a final prediction as shown in Figure 1. Our teacher model is trained in two stages. Firstly, it is pre-trained on large scale Wearable Stress and Affect Detection (WESAD) dataset [1] to distinguish between baseline and stress states of the driver by using the data from the chest-worn device of the WESAD dataset and the following physiological sensors: EMG (Electromyogram), ECG (Electrocardiogram), and GSR (Galvanic Skin Response) or Skin Conductance (SC). The resulting model is then fine-tuned on Stress Recognition in Automobile Drivers (SRAD) [2] dataset in order to further classify the stress states into low, medium, and high stress. The datasets for training the teacher are as follows:

1) **WESAD:** to discriminate between neutral and stress states we used a dataset for Wearable Stress and Affect Detection (WESAD) that contains data from 17 subjects obtained from two different devices: chest-worn (RespiBAN Professional) and wrist-worn (Empatica E4). A chest-worn device provides physiological signals of 6 modalities such as ECG, EDA, EMG, TEMP, and accelerometer (ACC) values in x, y, and z directions. All signals are sampled at 700 Hz. A wrist-worn device that provides data for 4 modalities BVP, EDA, TEMP, and ACC. This dataset contains labels for baseline, stress, and amusement.
2) **SRAD:** the resulting model is then further trained on the Stress Recognition in Automobile Drivers (SRAD) dataset to classify the recognized stress into low, medium, and high-stress states. The SRAD dataset contains physiological signals of 15 subjects in three different road scenarios i.e. rest, city driving, and highway driving, providing following sensors data: EMG, ECG, GSR, Heart Rate (HR), and Respiration (RESP).

2) The Student: Our student model is based on Temporal Convolutional Networks (TCN). The input to our TCN are the features extracted from video frames using the tool for example, OpenFace [10]. OpenFace is an open-source software that has been used to extract 68 key features of the face and used in emotion recognition analysis. The student model is trained from scratch and monitor its progress on the validation set and then select the final model that minimizes our cross-entropy loss over this validation set. The dataset that we use for this cross-modal process is BioVid Emo dB [11]. It is a multimodal database that contains physiological sensor data and corresponding videos to classify the emotions into 5 discrete states of 94 subjects. Out of these 5 emotions, three are of our interest including fear, anger, and disgust. The resulting student model is then deployed for inference on resource constrained devices.

IV. Conclusions

In this work, we proposed the methodology of detecting stress using a large-scale dataset for cross-modal distillation from sensors to videos. The benefits of this approach are evident because, during the inference of the final model, drivers do not need to wear the sensors as that could be annoyed for drivers and it will add noise to the sensor values. The other great benefit of this approach is that we can have a large number of unlabeled videos available to detect emotions and stress.

References

[1] Y. Liu, K. Wang, G. Li, and L. Lin, “Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition,” *IEEE Transactions on Image Processing*, 2021.
[2] K. Young, M. Regan, and M. Hammer, “Driver distraction: A review of the literature,” *Distracted driving*, vol. 2007, pp. 379–405, 2007.
[3] M. A. Tischler, C. Peter, M. Wimmer, and J. Voskamp, “Application of emotion recognition methods in automotive research,” in *Proceedings of the 2nd Workshop on Emotion and Computing—Current Research and Future Impact*, vol. 1, 2007, pp. 55–60.
[4] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 205–211.
[5] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
[6] L. J. Ba and R. Caruana, “Do deep nets really need to be deep?” *arXiv preprint arXiv:1312.6184*, 2013.
[7] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, “Learning small-size dnn with output-distribution-based criteria,” in *Fifteenth annual conference of the international speech communication association*, 2014.
[8] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, “Emotion recognition in speech using cross-modal transfer in the wild,” in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 292–301.
[9] S. Gupta, J. Hoffman, and J. Malik, “Cross modal distillation for supervision transfer,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.
[10] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.
[11] L. Zhang, S. Walter, X. Ma, P. Werner, A. Al-Hamadi, H. C. Traue, and S. Gruss, ““biovid emo db”: A multimodal database for emotion analyses validated by subjective ratings,” in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–6.

[1] https://archive.ics.uci.edu/ml/datasets/WESAD/
[2] https://physionet.org/content/drivedb/1.0.0/